UiO : Matematisk institutt
Det matematisk-naturvitenskapelige fakultet

KORRELASJON OG IKKE-PARAMETRISKE PROBLEMER

MOD5960 - Modellering og dataanalyse

Einar Christopher Wellén

Masteroppgave, våren 2015

Innhold

1 Introduksjon **6**

1.1 Bakgrunn for oppgaven 6
1.2 Korrelasjon 6
1.3 Variabeltype 7
1.3.1 Kvalitative variable 7
1.3.2 Kvantitative variable 8
1.4 Korrelasjonstyper og korrelasjonsindekser 8
1.5 Misledende/illusorisk korrelasjon mellom to variable. Partielle korrelasjoner. 11
1.6 Semipartielle korrelasjoner ved bruk av multippel linær regresjon. 11
1.7 Software 11
1.8 Veien videre 12

2 Pearson's r **14**

2.1 Definisjon Pearson's r 14
2.1.1 For en populasjon 14
2.1.2 For et utvalg(måling) 14
2.2 Generalisert korrelasjonskoeffisient 16
2.2.1 Definisjon 16
2.2.2 Pearson's r et spesialtilfelle av den generaliserte korrelasjonskoeffisienten 16
2.3 Matematiske egenskaper og interpretasjon 17
2.3.1 Sammenhengen mellom Pearson's r og minste kvadrater regresjonsanalyse 19
2.3.2 Geometrisk interpretasjon 20
2.4 Eksistens og avhengighet av fordelingsantagelser 20
2.5 Utvalgsstørrelse og inferens 20
2.5.1 Forutsetniger på Pearson's r ved statistisk inferens . . . 20
2.5.2 Normalfordelte data 21
2.6 Robusthet 29
2.7 To eller flere uavhengige utvalg 29
2.7.1 To uavhengige utvalg - H_0: $\rho_1 = \rho_2$ 29
2.7.2 Avhengige korrelasjoner - H_0: $\rho_1 = \rho_2$ 31
2.8 Spesialtilfeller av Pearson's r 33

2.8.1 Den biserielle punktkorrelasjonen mellom en binær og kontinuerlig variabel - r_{pb} 34
2.8.2 ϕ-koeffisienten for korrelasjon mellom to binære variable 35
2.8.3 Rådata kontra transformerte data 35
2.8.4 η-koeffisienten for korrelasjon mellom en multippel og en kontinuerlig variabel - η 37
2.8.5 Korrelasjon mellom en binær og multippel variabel - r_{MD} 39
2.8.6 Korrelasjon mellom en binær og en ordinal variabel - r_{DR} 39
2.8.7 Korrelasjon mellom en multippel og ordinal variabel - r_{MR} 39
2.8.8 Korrelasjon mellom en ordinal og en kontinuerlig variabel - r_{RI} 40

3 Permutasjonstester og bootstrapping - Pearson's r 40
3.1 Permutasjonstester 40
3.1.1 Eksakt permutasjonsfordeling 41
3.1.2 Tilpasning av den eksakte permutasjonsfordelingen til fordelingens momenter 42
3.1.3 Tilnærmet permutasjonstest 45
3.2 Bootstrapping 46
3.2.1 Bootstrapping og korrelasjonskoeffisienten ρ 48

4 Ikke-parametriske korrelasjonsmål 51
4.1 Spearman's rang korrelasjonskoeffisient r_s - et estimat for populasjonskorrelasjonskoeffisienten ρ_s 53
4.1.1 Spearman's r_s et spesialtilfelle av den generaliserte korrelasjonskoeffisienten 53
4.2 Kendall's rang korrelasjonskoeffisient t - et estimat for populasjonskorrelasjonskoeffisienten τ 54
4.2.1 Kendall's t sett på som en koeffisient av overensstemmelse(concordance) 54
4.2.2 Kendall's t sett på som en koeffisient av uorden(disarray) 56
4.2.3 Kendall's t - et spesialtilfelle av den generelle korrelasjonskoeffisienten 58
4.3 Like rangeringer (Tied ranks) 58
4.3.1 Beregning av Kendall's t ved like rangeringer 59
4.3.2 Beregning av r_s ved like rangeringer 61
4.3.3 Like rangeringer herunder en binær variabel 62
4.3.4 Like rangeringer med to binære variable 62

4.4 Signifikanstester for r_s og Kendall's t når vi ønsker å teste om variablene er uavhengige - korrelasjon lik 0 63
4.4.1 Signifikanstester Kendall's t 64
4.4.2 Signifikanstester r_s 68
4.5 Signifikanstester og konfidensintervaller 70
4.5.1 H_0: $\tau = \tau_q$ og H_0: $\rho_s = \rho_q$ 73
4.5.2 H_0: $\tau_q = \tau_p$ og H_0: $\rho_q = \rho_p$ 74
4.5.3 Inferensproblemer 75
4.6 Eksistens og avhengighet av fordelingsantagelser 75
4.7 Robusthet 75
4.8 Utvalgsstørrelse og inferens 75
4.9 Nærmere om valget mellom r, r_s og Kendall's t 76
4.10 Forholdet mellom varianser(effisiens) og uavhengighetstester . 77

5 Misvisende korrelasjon 78
5.1 Tredje variabler - partiell korrelasjon 79
5.1.1 Pearson's r og førsteordens partiell korrelasjon 81
5.1.2 Kendall's t og førsteordens partiell korrelasjon 83
5.2 Semipartielle korrelasjoner 83
5.3 Pearson's r og aggregerte målinger 85
5.4 Kort om restriksjoner på hvilke verdier en variabel kan ta . . . 85
5.5 Kort om målefeil 86

6 Analyse av data på Norges beste langrennsutøvere 86
6.1 Data 86
6.2 Aktuelle korrelasjonsmål og teknikker for å utføre inferens på data 87
6.3 Bruk av korrelasjonsmål i toppidretten 89
6.3.1 n liten 89
6.3.2 Linearitet eller monotonisitet. Årsakssammenheng. . . 90
6.4 Hypotesene 91
6.4.1 Hastighet sammenlignet med vekt(H_1), høyde(H_2) og bmi(H_3) - 2014 data. 92
6.4.2 Hastighet sammenlignet med laktatmaxDIA(H_7) og laktatmaxSTA(H_8) - 2014-data 95
6.4.3 Hastighet sammenlignet med VO2maxDIA(H_4), VO2maxSTA(H_5) og VO2maxSTA/VO2maxDIA(H_6) - 2014-data 95
6.4.4 Gjennomsnittshastighet på hele løpet sammenlignet med hastighet i de ulike segmenter(H_9) - 2014-data jfr. tabell 19 på side 101 100

6.4.5 Puls i de ulike segmenter og gjennomsnittspuls mot gjennomsnittshastighet totalt(H_{10}) - 2014-data jfr. tabell 20 på side 104 103
6.4.6 Hvor hardt du tar i av gjennomsnittspuls i de ulike segmentene mot gjennomsnittelig hastighet totalt(H_{11}) - 2014-data 105
6.4.7 Gjennomsnittshastighet på hele løpet sammenlignet med hastighet i de ulike segmenter (H_{12}) - 2013-data jfr. tabell 21 på side 107 106
6.4.8 Puls i de ulike segmenter og gjennomsnittspuls mot gjennomsnittelig hastighet(H_{13}) - 2013-data 109
6.4.9 Hvor hardt du tar i av gjennomsnittspuls i de ulike segmentene sammmenlignet med gjennomsnittelig hastighet(H_{14}) - 2013-data 110
6.4.10 Korrelasjoner fra to uavhengige utvalg 2013/2014 . . . 110
6.4.11 Partielle korrelasjoner - 2014 110

7 Sluttkommentarer **112**

A Appendiks 1: Metodene **117**

1 Introduksjon

1.1 Bakgrunn for oppgaven

Utgangspunktet for masteroppgaven var et samarbeidsprosjekt med Olympiatoppen og Norges Skiforbund langrenn under ledelse av Øyvind Sandbakk(Fag- og FOU ansvarlig ved Olympiatoppen Midt-Norge). Prosjektet ser blant annet ut til å resultere i artikkelen 'Speed and heart rate profiles in skating and classical crosscountry skiing competitions'. Jeg bidro med statistiske beregninger og endel av analysene ble gjort ved hjelp av diverse korrelasjonsmål. Tilgangen på lite data, og herunder stor usikkerhet knyttet til fordelingsegenskaper ved data, medførte at vi i stor grad tok i bruk fordelingsfrie statistiske prosedyrer/mål.
Ideén med masteroppgaven er å gå nærmere inn på bakgrunnen for ovenfornevnte korrelasjonsmål herunder undersøke egenskaper, som avhengighet av fordelingsantagelser, robusthet mot ulike fordelinger av data, og hvordan tilnærmelser avhenger av utvalgsstørrelse. Det vil spesielt bli aktuelt å gå nærmere inn på korrelasjonsmål, som er uavhengig av populasjonsfordelingen til variablene(X,Y) korrelasjonen beregnes for. Det blir herunder aktuelt å sammenligne ulike fordelingsfrie korrelasjonsmål med korrelasjonsmål avhengig av fordelingen på populasjonsfordelingen(spesielt hvor vi antar en bivariat normalfordeling på X,Y).

1.2 Korrelasjon

Et svært sentralt spørsmål i dataanalyse er hvorvidt det er en sammenheng mellom en variabel X og en variabel Y. En måte å besvare dette spørsmålet på er å beregne korrelasjonen mellom de to størrelsene. Korrelasjon måler assosiasjonen mellom to variable både med hensyn på styrke og ofte retning. Avhengig av hvilket korrelasjonsmål, som benyttes, er korrelasjon typisk et mål for den lineære sammenhengen mellom variablene X og Y jfr. Pearson's r omhandlet i 2 på side 14, eller et mål for den monotonistiske sammenhengen jfr. Kendalls t og Spearmsn's r_s omhandlet i 4 på side 51.
Hvis høye verdier av variabelen X er i par med høye verdier av variabelen Y er størrelsene X og Y positivt korrelerte. Hvis lave verdier av X er i par med høye verdier av Y er de to variablene negativt korrelerte. Hvis vi f.eks. korrelerer gjennomsnittshastigheten, fra den nasjonale åpningen på Beitostølen November 2014, for kvinner klassisk 10km, med de samme kvinnenes

beregnede VO2max(maksimalt oksygenopptak) får vi en klart positiv korrelasjon. Resultatet antyder at kvinnelige langrennsløpere i norgestoppen med høy VO2max presterer bedre enn kvinnelige langrennsløpere i norgestoppen med lavere VO2max.
Det er viktig å skille på korrelasjon og årsakssammenheng. Ofte er to variable statistisk relatert, men det eksisterer likevel ingen årsakssammenheng mellom dem. Det er sikkert mulig å finne en sammenheng, via korrelasjon, mellom antall benbrudd og bønn om ikke å bli skadet, men det eksisterer ingen årsakssammenheng mellom de to nevnte variablene. Det kan også være en tredje variabel, Z, som påvirker både X og Y på en slik måte at det oppstår korrelasjon mellom størrelsene X og Y og korrelasjonen mellom X og Y er da misledende jfr. 5 på side 78 om misvisende korrelasjon. Korrelasjon er kun et numerisk mål på om to størrelser varierer i takt. At to størrelser varierer i takt betyr ikke at det er en årsakssammenheng.
Det er videre viktig å skille mellom korrelasjon og regresjon. Regresjon er avhengighet mellom en avhengig variabel Y og en eller flere uavhengige variable $X_1, X_2, X_3, ..X_n$. Regresjon kalles å tilpasse en linje/kurve, plan/flate, eller hyperplan/hyperflate til dataene og kan f.eks. brukes til å forutsi/predikere en variabelverdi ut fra de uavhengige variablene. Ved bruk av korrelasjon er hverken X eller Y uavhengig variabel.
I oppgaven her ser jeg kun på korrelasjon mellom to variable og avgrenser dermed mot kanonisk korrelasjonsanalyse.

1.3 Variabeltype

1.3.1 Kvalitative variable

En variabel kan være kvalitativ og betegnes da enten som kategorisk(nominal), eller ordinal.
Kategoriske variable plasserer individer i kategorier, eller grupper, som f.eks kjønn(mann og kvinne), eller skimerke (fisher, madshus, rossignol og atomic). En kategorisk variabel er en faktor med to(binær), eller flere(multiple) nivåer. En variabel, som er kategorisert i to kategorier, sies å være dikotomisert og vil heretter bli kalt binær. En variabel, som er kategorisert i flere kategorier, sier jeg å være polytomisert og vil heretter bli kalt multippel. Kategoriske variable måles på nominal skala, som er en skala hvor målingene bare kategoriseres etter 'navn'(kategori). Eksempler på kategoriske data er tellinger(frekvenser) eller % (relative frekvenser), som faller i ulike kategorier. Data, som er klassifisert i flere kategorier, men som har en naturlig orden; f.eks karakterskalaen A, B, C, D, E, F, der vi f.eks setter 1=A, 2=B, 3=C, 4=D, 5=E og 6=F,

kalles ordinale eller rangerte data. Et annet eksempel på en ordinal variabel er Likertskalaen, som måler ulike holdninger hos respondenten, der f.eks. 1=meget uenig, 2=uenig, 3=nøytral, 4=enig, og 5=meget enig. Det er åpenbart at data klassifisert i flere kategorier og som har en naturlig orden er ordinale, men man kan også 'transformere' kvantitative data(jfr. 1.3.2) til ordinale data ved å rangere dataene. Ordinale data måles på en ordinal skala. Med ordinal skala forstås nummer som indikerer observasjonens relative posisjon(rang), men nummeret sier ingenting om styrken, eller forskjellen, mellom observasjonene annet enn avstanden i posisjon.

1.3.2 Kvantitative variable

En variabel kan være kvantitativ og betegnes da enten som kontinuerlig, eller diskret. Kvantitative variable får en numerisk verdi, som det utføres regneoperasjoner på. Kvantitative variable kan være kontinuerlige, som tid, hastighet, temperatur og VO2-opptak, eller de kan være diskrete(diskontinuerlige), som antall barn, øyne på en terning etc.. Diskrete fordelinger antar tellbare, eller endelig mulige verdier. Kvantitative variable måles på en intervallskala eller forholdstallskala. På en intervallskala gir det mening å tallfeste avstanden mellom verdiene, d.v.s det gir f.eks. mening å subtrahere verdiene. Intervallskala, som i tillegg har et absolutt nullpunkt(f.eks. alder, antall øyne og VO2max), kalles en forholdstallskala. På en forholdstallskala gir det mening å beregne forholdet mellom verdier(dobbelt antall øyne, 30% høyere VO2max etc.).
I oppgaven her forstås kontinuerlige data, som kvantitative data, dersom ikke noe annet fremgår av konteksten.

1.4 Korrelasjonstyper og korrelasjonsindekser

I henhold til de tre variabeltypene(nominal, kontinuerlig og ordinal) definert i forrige paragraf kan vi lage 10 mulige koblinger av relasjoner(korrelasjonstyper) mellom to variable. Videre er det utviklet flere korrelasjonsindekser for å estimere/måle de ovenfornevnte relasjoner. I tabell 1 på neste side jfr. [5] fremkommer 20, mye brukte, ulike korrelasjonsindekser relatert til de 10 mulige typer av relasjoner mellom to variable.
Det kanskje mest brukte korrelasjonsmål verden over er Pearson produktmoment korrelasjonskoeffisienten(r). Pearson's r kan brukes som deskriptiv statistikk på samme måte som gjennomsnitt, typetall, eller standardavvik.
Ofte ønsker vi, når vi beregner den empiriske korrelasjonskoeffisienten Pear-

Tabell 1: Korrelasjonsindekser

	Kvalitative variable			Kvantitative variable
	Kategoriske/nominale			
Variabeltype	Binær	Multippel	Ordinal	Kontinuerlig og diskret
Binær	$\phi^{1,2}{}_{\text{(phi)}}$ $r_{tet}{}_{\text{(tetrachoric)}}$			
Multippel	r^1_{MD}	$C^2{}_{\text{(contingency)}}$ $V^2{}_{\text{(Cramer)}}$		
Ordinal	r^1_{DR}	r^1_{MR}	$r_s^{1,2}{}_{\text{(Spearman)}}$ $t^2{}_{\text{(Kendall)}}$ $t_a^2{}_{\text{(Kendall)}}$ $t_b^2{}_{\text{(Kendall)}}$ $t_c^2{}_{\text{(Stuart og Kendall)}}$ $\gamma^2{}_{\text{(Goodman og Kruskal)}}$ $t^2_{xy,z}{}_{\text{(Kendall partial rank-order)}}$	
Kontinuerlig, og diskret	$r^1_{pb}{}_{\text{(point-biserial)}}$ $r_{bis}{}_{\text{(biserial)}}$	$\eta^1{}_{(eta)}$	r^1_{RI}	$\mathrm{r}_{\text{(Pearson)}}$ $r^1_{xy,z}{}_{\text{(førsteordens partiell)}}$
1 = Spesialtilfelle av Pearson r, 2 = Ikke - parametrisk korrelasjonskoeffisient				

son's r på bakgrunn av det utvalget(målingen n) vi har, å estimere den teoretiske korrelasjonskoeffisienten ρ for populasjonen vi studerer. Heretter defineres ρ, som populasjonskorrelasjonskoeffisienten for Pearson's r. Fordelingen, til et utvalg fra populasjonen, vil jeg i oppgaven betegne, som målefordelingen, eller alternativt utvalgsfordelingen.
Korrelasjonskoeffisienter, brukt på ikke-parametriske problemstillinger(ikke-parametriske korrelasjonsmål jfr. 4 på side 51), krever generelt færre forutsetninger. I motsetning til ved Pearson's r vil vi f.eks., ved inferens knyttet til en ikke-parametrisk korrelasjonskoeffisient, ikke kreve at (X,Y) kommer fra en bivariat normalfordeling.
Videre eksisterer ikke-parametriske korrelasjonsmål, som måler assosiasjon(styrke, men ikke retning) mellom kategoriske(nominale) variable etter å ha systematisert dataene i krysstabeller. Krysstabellene viser nivåene(kategoriene) og tellingene for alle kombinasjoner av nivåer for de kategoriske variablene. Jeg må dessverre avgrense oppgaven mot utledning og praktisk bruk av disse korrelasjonsmål grunnet plassmangel.
Blant de 20 korrelasjonsindeksene i tabell 1 på forrige side er det 8 spesialtilfeller av Pearson's r: 1) den ikke-parametriske phi-koeffisienten (ϕ) for to binære variable, 2) den parametriske eta-koeffisienten (η) for en multippel og en kontinuerlig variabel, 3) den ikke-parametriske Spearman-koeffisienten(r_s) for to ordinale variable, 4) den parametriske punkt-biserielle korrelasjonskoeffisienten (r_{pb}) for en binær og en kontinuerlig variabel, og de fire mindre kjente parametriske korrelasjonskoeffisientene r_{MD} for en binær og multippel variabel, r_{DR} for en binær og en ordinal variabel, r_{MR} for en multippel og en ordinal variabel, og r_{RI} for en ordinal og kontinuerlig variabel. Den 'opprinnelige' Pearson's r gjelder for to kontinuerlige variable.
Dersom en av to kontinuerlige variable 'transformeres' til en binær variabel kan vi beregne korrelasjonen mellom den nye binære og den kontinuerlige variabelen ved hjelp av korrelasjonsmålestimatet, for Pearsons r, r_{bis}. Dersom to kontinuerlige variable 'transformeres' til to binære variable kan vi beregne korrelasjonen mellom de to 'nye' binære variablene ved hjelp av korrelasjonsestimatet,for Pearson's r, r_{tet}.
I tillegg til de to ikke-parametriske korrelasjonskoeffisientene(ϕ og r_s), nevnt ovenfor, fremkommer det av tabell 1 på forrige side, ytterligere 7 ikke-parametriske korrelasjonskoeffisienter: Contingency-koeffisienten C og Cramer's V-koeffisient, som begge gjelder for to multiple kategoriske variable, Kendall's t-koeffisienter(t, t_a og t_b), Kendall-Stuart's t_c-koeffisient og Goodman and Kruskal's γ-koeffisient, som alle gjelder for to ordinale variable.

1.5 Misledende/illusorisk korrelasjon mellom to variable. Partielle korrelasjoner.

Dersom vi antar at det eksisterer en misledende, eller illusorisk, sammenheng mellom to variable, som følge av påvirkning fra en eller flere andre variable kan vi teste dette ved å bruke ulike statistiske kontrollprosedyrer jfr. 5 på side 78. Hvis vi ønsker å kontrollere kun for en enkelt tredje variabel kan vi benytte oss av en førsteordens partiellkorrelasjon. En førsteordens partiellkorrelasjon refererer til assosiasjonen mellom variablene X og Y etter at effekten av den tredje variabelen Z er fjernet(statistisk kontrollert for). Under antagelsen om at Z er den eneste variabelen, som påvirker variablene X og Y, og at Z ikke påvirkes av verken X eller Y er den førsteordens partielle korrelasjonen den faktiske korrelasjonen, og relasjonen mellom X og Y, uten å kontrollere for Z, anses å være illusorisk.

1.6 Semipartielle korrelasjoner ved bruk av multippel linær regresjon.

Bruk av semipartielle korrelasjoner, jfr. 5.2 på side 83, er en måte å angi den relative viktigheten av uavhengige variable når man bestemmer den avhengige variabelen Y ved multippel lineær regresjon. I hovedsak viser de semipartielle korrelasjonene hvor mye hver variabel unikt bidrar til den multiple korrelasjonskoeffisienten R^2 smln. med de resterende variablene, som inngår i regresjonen. Multippel lineær regresjon krever, etter mitt skjønn, blant annet god innsikt i korrelasjonsanalyse herunder semipartielle og partielle korrelasjoner, for å bygge en best mulig modell.
Jeg vil i oppgaven her studere semipartielle korrelasjoner da de f.eks. er nyttige i analysen av hvorvidt gjennomsnittshastighetene i de ulike segmentene av løypa(oppover, bortover og nedover) påvirker gjennomsnittshastigheten på hele distansen ulikt eller ikke.

1.7 Software

Statistiske analyser er gjort i R(R Development Core Team, Vienna, Austria). For mer informasjon se hjemmesiden til 'The R Project for Statistical Computing': http://www.r-project.org/.
Jeg har benytttet meg av pakker og funksjoner i R jfr. tabell 2 på side 13. Jeg har ved hjelp av nevnte pakker og funksjoner og noe annen implementert kode laget 12 ulike metodefunksjoner, jfr. Appendix A på side 117. De tolv metodefunksjonene er implementert i filen 'Vedlegg1ProgrammeringsfilMasteroppgaveEinarChristopherWellén.R' og ved-

lagt oppgaven.
Kode for beregningene(analysene), som er blitt utført i R, er implementert i filen 'Vedlegg2AnalysefilMasteroppgaveEinarChristopherWellén.R' og vedlagt oppgaven.
Jeg har også ved enkelte anledninger benyttet IBM SPSS Statistics versjon 22, som en kontroll ved beregninger gjort i R.

1.8 Veien videre

Jeg vil i seksjon 2 på side 14 se nærmere på det parametriske korrelasjonsmålet Pearson's r. Vi viser hvordan Pearsons'r er utledet fra det generelle korrelasjonsmålet Γ og studerer spesifikt interpretasjon og matematiske egenskaper, eksistens og avhengighet av fordelingsantagelser, hvordan tilnærminger avhenger av utvalgsstørrelse og robusthet. Jeg ser så på hvordan man kan teste om to korrelasjoner beregnet på to uavhengige utvalg, eller på to avhengige utvalg er signifikant forskjellige. Tilslutt ser jeg på variasjoner og bruk av Pearson's r når vi ikke har to kvantitative variable.
I seksjon 3 på side 40 ser vi kort på bruk av de fordelingsfrie metodene permutasjonstester og bootstrappingstester relatert til Pearson's r.
I seksjon 4 på side 51 studerer jeg de ikke-parametriske korrelasjonskoeffisientene Spearman's r_s, Kendall's t, Kendall's t_a og Kendall's t_b. Jeg viser blant annet hvordan Kendall's t og r_s er utledet fra den generelle korrelasjonskoeffisienten Γ. Vi ser også på hvordan man behandler likt rangerte data. Deretter ser vi på signifikanstester, konfidensintervaller, eksistens og avhengighet av fordelingsantagelser, robusthet, og hvordan tilnærminger avhenger av utvalgsstørrelse for de ikke-parametriske korrelasjonsmålene Kendall's t og r_s. Til slutt diskuterer jeg valget mellom de tre korrelasjonskoeffisientene Spearman's r_s, Pearson's r og Kendall's t og ser på de asymptotiske relative effisiensene(ARE) for permutasjonstesten til r, Kendall's t-testen og r_s-testen for uavhengighet relativt til den ordinære målekorrelasjonskoeffisienten Pearson's r når den alternative hypotesen er at X,Y kommer fra en bivariat normalfordeling der ρ=0.
I seksjon 5 på side 78 om misvisende korrelasjon ser jeg nærmere på hvordan vi kan kontrollere for en tredje variabel Z ved beregning av korrelasjonen mellom to variabler X og Y(konfundering). Jeg vil spesielt se nærmere på de to første-ordens partielle korrelasjonskoeffisientene $r_{XY,Z}$ og $t_{XY,Z}$ for henholdsvis to kontinuerlige variable og to ordinale variable jfr. tabell 1 på side 9. Videre vil jeg se på hvordan semipartielle korrelasjoner kan benyttes til interpretasjon av hvordan flere variable påvirker hverandre og spesielt vil jeg se på dette i en regresjonskontekst. Til slutt ser vi kort på andre forhold(aggregerte målinger, verdirestriksjoner og målefeil), som kan medføre at beregnede kor-

Tabell 2: Funksjoner brukt i R

Pakke	Funksjon	Bruk
base	length,tanh,log,abs, sqrt,factorial,do.call, rbind,rep,ceiling, floor,as.integer,sum, sample,mean,sort	
stats	pnorm,qnorm,pt,qt	Beregning av p-verdier og kvantiler
stats	rnorm,sd	Trekk av tall fra normalfordelingen, standardavvik
stats	cor	Beregning av Pearson's r korrelasjon
Kendall	Kendall	For Kendall's t, t_a, t_b og score S(ved kontinuitetskorreksjon)
pspearman	spearman.test	Brukes ved beregning av Spearman's r
SuppDists	pPearson,qPearson pKendall,qKendall pSpearman,qSpearman	Beregning av p-verdier og kvantiler
combinat	permn	Genererer n! permuteringer av n elementer
DescTools	GoodmanKruskalGamma	Gir gammakoeffisienten
ppcor	pcor.test,pcor	Partielle korrelasjoner ved kontroll av en(pcor.test), eller flere(pcor) variable
ppcor	spcor.test, spcor	Semipartielle korrelasjoer
psych	r.test	Teste forskjell på avhengige korrelasjoner, som 1)deler en variabel(William's test), eller 2) involverer ulike variable(Steiger test)
stats	shapiro.test	Tester om data kan anses normalfordelt univariat
mvnormtest	mshapiro.test	Tester om data kan anses normalfordelt multivariat
stats	bartlett.test	Tester homogenitet av varianser i to utvalg.
stats	t.test	En- og toutvalgs t-tester. Toutvalgs t-tester kan være parret(målinger på samme populasjon), eller uparret(målinger på uavhengige populasjoner).
stats	wilcox.test	Toutvalgs ikke-parametrisk test. Parret(wilcoxson signed rank test) Uparret(Mann Whitney U/Wilcoxon rank sum) Testene tilsvarer de parametriske t-testene.

relasjoner er misvisende.
I seksjon 6 på side 86 tester jeg utvalgte hypoteser knyttet til data fått fra Olympiatoppen og Norges Skiforbund langrenn. Jeg beskriver først dataene og valg og bruk av korrelasjonsmål i toppidretten. Deretter defineres og testes hypotesene ved hjelp av de parametriske og ikke-parametriske korrelasjonsmålene redegjort for i de foregående kapitlene.
Avslutningsvis, i seksjon 7 på side 112, oppsummerer, og diskuterer jeg anvendelse og interpretasjon av korrelasjonsmål.

2 Pearson's r

Pearson's produkt-moment korrelasjons koeffisient r, utviklet av Karl Pearson fra en tidligere idé utviklet av Francis Galton i 1880-årene, er antagelig den mest brukte observatoren for å måle relasjon mellom variable. Det har blitt estimert at Pearson's r og dets spesialtilfeller blir valgt i 95% av tiden i forskning, som går med på å beskrive en sammenheng mellom to størrelser, eller til å gjøre inferens relatert til populasjonskorrelasjonen ρ jfr. [6].

2.1 Definisjon Pearson's r

2.1.1 For en populasjon

Pearson's korrelasjonskoeffisient, som er en bivariat stokastisk variabel, er for hele populasjonen gitt ved:

$$\rho_{X,Y} = \frac{COV(X,Y)}{\sigma_X * \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X * \sigma_Y} \tag{1}$$

,der COV er kovariansen mellom X og Y, σ_X og σ_Y er standardavviket for henholdsvis X og Y, og μ_X og μ_Y er forventningsverdien for henholdsvis X og Y. Siden $\mu_X = \mathrm{E(X)}$, $\sigma_X^2 = E[(X-E(X))^2] = E(X^2) - E^2(X)$ og tilsvarende for Y og siden $E[(X-E(X))(Y-E(Y))] = E(XY) - E(X)E(Y)$ kan vi skrive korrelasjonskoeffisienten, som:

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}} \tag{2}$$

2.1.2 For et utvalg(måling)

Pearson's korrelasjonskoeffisient for et utvalg er gitt ved:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{3}$$

$$\Updownarrow$$

$$r = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{X_i - \bar{X}}{s_X})(\frac{Y_i - \bar{Y}}{s_Y}) \tag{4}$$

$$\Updownarrow$$

$$r = \frac{1}{n-1}\sum_{i=1}^{n} Z_{X_i} Z_{Y_i} \tag{5}$$

,der $Z_{X_i} = \frac{X_i - \bar{X}}{s_X}$, $Z_{Y_i} = \frac{Y_i - \bar{Y}}{s_Y}$, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ og $s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$, $s_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$ er henholdsvis standard score, målegjennomsnitt og målestandardavvik for respektive X og Y.
Pearson's r kalles produkt-moment korrelasjon fordi den blir kalkulert ved å multiplisere standardscoren for to variable(produkt av to variable) for så å beregne gjennomsnittet(momentet) for disse produktene basert på et gitt antall målinger n. Jeg velger her å dele på n-1 og ikke n for at standardavvikene, som inngår i definisjonene 4 og 5 av r, skal være forventningsrette. Verdien til r endres ikke om vi deler på n istedenfor n-1 i definisjonene 4 og 5 så lenge vi deler på n og ikke n-1 ved beregning av s_X og s_Y. Definisjon 5 viser tydelig at vi kan oppfatte Pearson's r som en standardisert kovarians.
Vi kan også skrive Pearson's r på formen:

$$r = \frac{\sum_{i=1}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}(Y_i - \bar{Y})^2}} = \frac{\sum_i X_i Y_i - n\bar{X}\bar{Y}}{(n-1)s_X s_Y} \tag{6}$$

$$\Updownarrow$$

$$r = \frac{\sum_i X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\sum_i (X_i^2 - X_i\bar{X} - X_i\bar{X} + \bar{X}^2)}\sqrt{\sum_i (Y_i^2 - Y_i\bar{Y} - Y_i\bar{Y} + \bar{Y}^2)}} \tag{7}$$

$$\Updownarrow *\frac{n}{n}$$

$$r = \frac{n\sum_i X_i Y_i - \sum_i X_i \sum_i Y_i}{\sqrt{n\sum_i X_i^2 - (\sum_i X_i)^2}\sqrt{n\sum_i Y_i^2 - (\sum_i Y_i)^2}} \tag{8}$$

Vi ser at formelen kan løses numerisk uten å gjennom noen løkke, men den kan være ustabil avhengig av hvilke tall som er involvert.

2.2 Generalisert korrelasjonskoeffisient

2.2.1 Definisjon

Vi definerer den generaliserte korrelasjonskoeffisienten Γ jfr. [1] ved:

$$\Gamma = \frac{\sum a_{ij} b_{ij}}{\sqrt{\sum a_{ij}^2 \sum b_{ij}^2}} \tag{9}$$

Vi tar utgangspunkt i n parvise målinger av variablene X og Y. Vi nummerer så målingene fra 1 til n for å kunne identifisere en hvilken som helst rekkefølge. Observasjonene av variablene X og Y kan nå altså defineres med $x_1, x_2, ...x_n$ og $y_1, y_2, ..., y_n$. Verdiene er enten verdiene slik de observeres på en kontinuerlig skala, eller rangeringer. Til hvert enkelt par av observasjoner, f.eks den i'te(x_i, y_i) og j'te(x_j, y_j), gir vi en X-score merket a_{ij} slik at $a_{ij} = -a_{ji}$ og en Y-score kalt b_{ij}, der $b_{ij} = -b_{ji}$. Dersom i = j lar vi $a_{ij} = 0$. Γ-korrelasjonen har egenskapene:
1) Tar verdier i intervallet f.o.m -1 t.o.m 1 jfr. Cauchy-Schwartz ulikheten, som sier at:

$$(\sum ab)^2 \leq \sum a^2 \sum b^2 \tag{10}$$

2) Hvis to korresponderende par av rangeringer i de to rangeringene av X og Y ikke har samme rekkefølge(f.eks i<j for X og i>j for Y) og man så bytter rekkefølgen i et par så vil Γ øke forutsatt at scorene til a_{ij} og b_{ij} ikke er null og at scorene ikke synker med økende avvik på rangeringene. For bevis se [1, s. 28].
Korrelasjonskoeffisientene t, r_s og r oppfyller disse forutsetningene.

2.2.2 Pearson's r et spesialtilfelle av den generaliserte korrelasjonskoeffisienten

Jeg viser til [1] og setter $a_{ij} = x_j - x_i$ og $b_{ij} = y_j - y_i$. Videre har vi at:

$$\sum_{i,j}(x_j - x_i)(y_j - y_i) = \sum_{i,j}(x_j y_j - x_j y_i - x_i y_j + x_i y_i) \tag{11}$$

$$\Updownarrow$$

$$2n\sum_i x_i y_i - 2\sum_{i,j} x_i y_j = 2n\sum_i x_i y_i - 2\sum_i x_i \sum_j y_j = 2nCOV(X,Y) \tag{12}$$

og

$$\sum_{i,j}(x_j - x_i)^2 = \sum_{i,j}(x_j^2 - 2x_j x_i + x_i^2) = 2n\sum_i x_i^2 - 2(\sum_i x_i)^2 = 2nVAR(X) \tag{13}$$

$$\sum_{i,j}(y_j - y_i)^2 = \sum_{i,j}(y_j^2 - 2y_j y_i + y_i^2) = 2n\sum_i y_i^2 - 2(\sum_i y_i)^2 = 2nVAR(Y) \tag{14}$$

Dermed har vi at den generelle korrelasjonskoeffisienten Γ kan skrives som den ordinære produkt-moment korrelasjonen til X og Y:

$$\Gamma = \frac{\sum_{i,j} a_{ij}b_{ij}}{\sqrt{\sum_{i,j} a_{ij}^2 \sum_{i,j} b_{ij}^2}} = \frac{COV(X,Y)}{\sqrt{VAR(X)VAR(Y)}} \tag{15}$$

2.3 Matematiske egenskaper og interpretasjon

Vi har følgende egenskaper for Pearson's r jfr. [5]:

- Pearson'r innehar egenskapene til det generelle korrelasjonsmålet Γ jfr. seksjon 2.2 på forrige side.
- Pearson's r lik 1 eller -1 korresponderer, for utvalgsfordelingen, til datapunkter liggende eksakt på en rett linje og for en bivariat fordeling vil hele fordelingen ligge på en rett linje dersom korrelasjonen ρ er lik 1 eller -1.
- Både variabel X og Y blir behandlet symmetrisk slik at korrelasjonen r mellom X og Y er den samme, som mellom Y og X.
- Absoluttverdien til Pearson's r blir ikke påvirket av lineære transformasjoner på X og Y og er m.a.o invariant under endringer i lokalisering og skalering. Dersom $A_i = cX_i + d$ og $B_i = eY_i + f$, der c$\neq$ 0, $e \neq 0$ og d, f er konstanter, kan vi matematisk vise at ved standardisering $|Z_{X_i}| = |Z_{A_i}|$ og $|Z_{Y_i}| = |Z_{B_i}|$ og dermed at $|\frac{1}{n-1}\sum_{i=1}^n Z_{X_i}Z_{Y_i}| = |\frac{1}{n-1}\sum_{i=1}^n Z_{A_i}Z_{B_i}| = |r|$ smln. likning 5 på side 15.
- Hvis X og Y ikke er innbyrdes avhengige(uavhengige) impliserer dette at populasjonskorrelasjonskoeffisienten (ρ) er lik 0. Dersom ρ=0 impliserer dette likevel ikke at X og Y nødvendigvis er uavhengige med mindre den bivariate normalantagelsen diskutert ovenfor holder - f.eks

vil en periodisk kurvet sammenheng mellom X og Y innebære at det er en innbyrdes sammenheng mellom X og Y, selvom Pearson's r=0. Det er nettopp her vanskeligheten ligger i å interpretere ρ, som en koeffisient for assosiasjon/avhengighet mellom variable generelt. Som vi så i foregående punkt er ρ essensielt bare en koeffisient for lineær assosiasjon/avhengighet mellom variable og mere komplekse former for sammenhenger/assosiasjon/avhengighet ligger utenfor dets rekkevidde å forklare. F.eks. vil en sammenheng mellom X og Y liggende på en helt symmetrisk parabel gi korrelasjon r=0 selvom det åpenbart er en klar sammenheng mellom endringen i X smln. med endringen i Y.
Generelt er det altså for vanskelig å samle simultanvariasjon i en enkelt koeffisient og jeg understreker at i praktisk arbeid anbefales det kun å bruke ρ, som et mål på assosiasjon/avhengighet i tilfeller hvor vi har normal, eller nær normal variasjon jfr. [3, p. 301] og [12, p. xxxv].

- Formen på fordelingene til X og Y påvirker muligheten for at Pearson's r kan oppnå maksimumsverdien 1 eller minimumsverdien -1. Dersom fordelingene til X og Y er symmetriske og har samme form(f.eks uniforme fordelinger, normalfordelinger, U-formede fordelinger) vil de mulige verdier for Pearson's r ligge i intervallet f.o.m. -1 t.o.m. 1. Maksimumsverdien til |r| vil være mindre enn 1 dersom formen på fordelingene til X og Y er ulik(f.eks der den ene er normalfordelt og den andre skjevfordelt). Dette er årsaken til at f.eks $|r_{pb}|$ og $|\phi|$ tenderer mot å være mindre enn 1. Gitt at begge fordelingene ikke deler samme form vil en økning i X ikke alltid bli fulgt av en økning i Y ved positiv sammenheng og ei heller vil en økning i X alltid bli parret med en nedgang i Y i tilfelle av negativ sammenheng. Jo mindre like formen på fordelingene til X og Y er jo mindre blir maksimumsverdien til |r|.

- Pearson's r kan ikke ses på, som et forholdstall, eller proporsjonal grad av sammenheng mellom X og Y slik at noe som har dobbelt så høy korrelasjon innebærer dobbelt så høy sammenheng mellom to størrelser. Derimot kan størrelen r^2 forstås som et forholdstall. r^2 forklarer andelen av variasjonen(fluktureringen) av en variabel(f.eks. Y), som er forklart ved en annen variabel(f.eks. X) og er m.a.o. forholdet mellom forklart variasjon og totalvariasjon. F.eks, hvis r = 0.97 så vil $r^2 = 0.9409$, som igjen betyr at 94.09% av variasjonen i y kan bli forklart av den lineære sammenhengen mellom X og Y(beskrevet ved hjelp av regresjonslinjen $\hat{Y} = aX+b$, der $\hat{Y}$ blant annet gir de tilpassede verdiene). De resterende 5,91% av totalvariasjonen i y forblir uforklart. Det er åpenbart at

$$0 \leq r^2 \leq 1.$$

For Pearson's koeffisient r = ± 1 impliserer dette linearitet, men for de ikke-parametriske rang-koeffisientene jeg introduserer nedenfor vil verdiene ± 1 vanligvis ikke implisere linearitet for rangene basert på underliggende kontinuerlige data. Ved bruk av rang-korrelasjoner hvor underliggende data er kontinuerlige er vi derimot interessert i det vi definerer som monotonisiteten til de kontinuerlige dataene. Hvis X og Y øker sammen er det en monotont økende sammenheng mellom X og Y. Dersom Y synker og X øker er relasjonen monotonisk synkende. For rang-korrelasjoner impliserer verdien 1 strengt økende monotonisitet og verdien -1 strengt synkende monotonisitet og altså ikke nødvendigvis full linearitet mellom de kontinuerlige variablene X og Y.

2.3.1 Sammenhengen mellom Pearson's r og minste kvadrater regresjonsanalyse

Både kvadratet av Pearson's $r_{X,Y}$ og koeffisienten R^2(andel av variansen i Y forklart ved de tilpassede verdiene $\hat{Y}$ via X) beregnet ved simpel lineær regresjon uttrykker det samme. Det er lett å vise denne sammenhengen. Ved å dekomponere totalvariasjonen til Y rundt sitt gjennomnitt får vi:

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{Y})^2 \tag{16}$$

$$\Updownarrow$$

$$1 = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{\sum_i (Y_i - \bar{Y})^2} + \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2} \tag{17}$$

, der $\hat{Y}_i$ er de tilpassede verdiene fra regresjonsanalysen, $\bar{Y}_i = \sum_i (Y_i/n)$ og n er antall observasjoner. Vi konstaterer at $\sum_i (\hat{Y}_i - \bar{Y})^2$ forklarer variansen i Y forklart av de tilpassede verdiene via X og at $\sum_i (Y_i - \hat{Y}_i)^2$ er den andelen av variasjonen i Y, som ikke kan forklares av X.
Beregner vi så kvadratet av Pearson's r for variablene Y og $\hat{Y}$ og benytter oss av at målekovariansen mellom $Y_i - \hat{Y}_i$ og $\hat{Y}_i$ er 0 får vi:

$$r^2_{Y,\hat{Y}} = \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2} = R^2 \tag{18}$$

Herav følger at $R^2 = r^2_{Y,X}$ siden Pearson's r ikke påvirkes av lineære transformasjoner på X jfr. 2.3 på side 17.
Ved multippel lineær regresjon(altså mer enn en forklaringsvariabel) kalles R^2 for den multiple korrelasjonskoeffisienten og vi har, grunnet mer enn to variable, ikke lenger en sammenheng mellom Pearsons r og R^2.

2.3.2 Geometrisk interpretasjon

For sentrerte data(altså data som er blitt fratukket gjennomsnittet slik at forventningen er 0) kan man vise at det er en sammenheng mellom korrelasjonskoeffisienten for de sentrerte dataene X^* til X og Y^* til Y og vinkelen θ^* mellom de to vektorene av sentrerte målinger. Korrelasjonskoeffisienten kan finnes ved $r_{X^*,Y^*} = cos(\theta^*) = \frac{\mathbf{x}^*\mathbf{y}^*}{||\mathbf{x}^*||||\mathbf{y}^*||}$
Dersom man bruker formelen for de sentrerte dataene på de usentrerte dataene vil vi få en annen verdi på korrelasjonskoeffisienten (dersom X og Y har ulike gjennomsnitt), men fortegnet vil ikke endre seg.
Jeg bemerker at det er etter å ha sentrert data at vi får en korrelasjonskoeffisient, som tilsvarer Pearson's r og at en geometrisk interpretasjon av Pearson's r krever at vi sentrerer data før vi finner korrelasjonskoeffisienten $cos(\theta^*)$.

2.4 Eksistens og avhengighet av fordelingsantagelser

Pearson's korrelasjonskoeffisient for en populasjon er definert i termer av momenter og eksisterer derfor for en hvilken som helst bivariat sannsynlighetsfordeling så lenge kovariansen for populasjonen er definert og variansene for marginalpopulasjonene er definert og ikke er 0. Cauchy-fordelingen har ingen endelig forventning og dermed udefinert varians og det følger at ρ ikke er definert for variabler X og Y, som følger slike fordelinger. I andre tilfeller med data, som antas å følge en tunghalet fordeling, vil problemer med endelig varians oppstå, men vi kan omgå dette hvis vi kan begrense verdiene variablene for slike tunghalede fordelinger kan ta!

2.5 Utvalgsstørrelse og inferens

2.5.1 Forutsetniger på Pearson's r ved statistisk inferens

Når Pearson's r blir beregnet for å beskrive karakteristikker for en måling krever den **ingen** forutsetninger. Derimot hvis man ønsker å bruke Pearson's r til å lage konfidensintervaller for populasjonskorrelasjonskoeffisienten ρ, eller utføre hypotesetester så antar man vanligvis at variablene X og Y kommer fra en bivariat normalfordeling. Selv når målinger$((x_1, y_1), (x_2, y_2), ..., (x_n, y_n))$ antas å komme fra en bivariat normalfordeling er det mer komplisert å gjøre inferens rundt en ukjent korrelasjonskoeffisient ρ basert på målekoeffisienten r enn inferens rundt forventning og varians til den enkelte fordeling X eller Y. Hvis (X,Y) ikke kommer fra en bivariat normalfordeling er parametrisk inferens rundt ρ enda vanskeligere.

2.5.2 Normalfordelte data

I en bivariat normalfordeling kan relasjonen mellom X og Y kun beskrives som en lineær kombinasjon jfr. [7]. Variablene X og Y følger en bivariat normalfordeling hvis og bare hvis det for hver mulige lineærkombinasjon av $Z = a_1X + a_2Y$ er slik at Z er normalfordelt og der konstantene a_1 og a_2 ikke er null. Det kan vises at marginalfordelingene til henholdsvis X og Y er normale. Videre er de betingede fordelingene Y|X(for gitte verdier av X) eller X|Y(for gitte verdier av Y) normalfordelte. Jeg bemerker at selvom marginalfordelingene til X og Y begge er normale ikke impliserer at den bivariate fordelingen til X,Y er normalfordelt da det fint går an å ha en ikke-lineær sammenheng mellom X og Y når begge marginalfordelingene er normale.
Hvis ρ=0 for en **bivariat normalfordeling** impliserer dette at X og Y er uavhengige, og dersom vår beregnede r er nær null støtter dette antagelsen om at $\rho = 0$.

Eksakt målefordeling når data følger en bivariat normalfordeling
Den simultane sannsynlighetsfordelingen til n målinger$(x_1, y_1), (x_2, y_2), .., (x_n, y_n)$ fra en bivariat normalfordeling med forventningsverdier μ_1 og μ_2 er gitt ved:

$$dF = \frac{1}{(2\pi)^n \sigma_1^n \sigma_2^n (1-\rho^2)^{\frac{1}{2}n}} e^{(-\frac{1}{2(1-\rho^2)}[\sum_i (\frac{x_i-\mu_1}{\sigma_1})^2 - 2\rho \sum_i \frac{(x_i-\mu_1)(y_i-\mu_2)}{\sigma_1\sigma_2} + \sum_i (\frac{y_i-\mu_2}{\sigma_2})^2])} dx_1 dy_1 dx_2 dy_2 ... dx_n dy_n \tag{19}$$

jfr. [2, s. 383 eq. (16.47)].
Man kan vise at uttrykket over kan uttrykkes ved hjelp av de fem parameterne til fordelingen(μ_1, μ_2, σ_1^2, σ_2^2, ρ) og korresponderende observatorer $\bar{x} = \frac{1}{n}\sum_i x_i$, $\bar{y} = \frac{1}{n}\sum_i y_i$, $s_1^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$, $s_2^2 = \frac{1}{n}\sum_i (y_i - \bar{y})^2$ og $r = \frac{1}{n}\sum_i \frac{(x_i-\bar{x})(y_i-\bar{y})}{s_1 s_2}$. Ronald Aylmer Fisher viste i 1915 [2, s. 385] jfr.[9] at det simultane frekvenselementet i likning 19, til de fem nevnte observatorene, er proporsjonal($\propto$) med:

$$e^{(-\frac{n}{2(1-\rho^2)}[[\frac{(\bar{x}-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(\bar{x}-\mu_1)(\bar{y}-\mu_2)}{\sigma_1\sigma_2} + \frac{(\bar{y}-\mu_2)^2}{\sigma_2^2}] + \frac{s_1^2}{\sigma_1^2} - 2\rho r\frac{s_1 s_2}{\sigma_1\sigma_2} + \frac{s_2^2}{\sigma_2^2}])} dv \tag{20}$$

, der dv $\propto s_1^{n-2} s_2^{n-2} ds_1 ds_2 (1-r^2)^{\frac{1}{2}(n-4)} dr d\bar{x} d\bar{y}$
Vi ser at uttrykket over kan faktoriseres i to deler slik at det ene uttrykket kun inneholder $\bar{x}$ og $\bar{y}$ og det andre uttrykket bare inneholder s_1, s_2 og r:

$$dF \propto e^{(-\frac{n}{2(1-\rho^2)}[\frac{(\bar{x}-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(\bar{x}-\mu_1)(\bar{y}-\mu_2)}{\sigma_1\sigma_2} + \frac{(\bar{y}-\mu_2)^2}{\sigma_2^2}])} d\bar{x} d\bar{y} \tag{21}$$

og

$$dF \propto e^{(-\frac{n}{2(1-\rho^2)}[\frac{s_1^2}{\sigma_1^2} - 2\rho r\frac{s_1 s_2}{\sigma_1\sigma_2 +} + \frac{s_2^2}{\sigma_2^2}])} s_1^{n-2} s_2^{n-2} (1-r^2)^{\frac{1}{2}(n-4)} ds_1 ds_2 dr \tag{22}$$

jfr. [2, s. 385], slik at i normale bivariate målefordelinger er fordelingen til gjennomsnittene helt uavhengig av fordelingen til variansene og kovariansen. Denne egenskapen er en karakteristisk egenskap ved alle multivariate normalfordelinger.
Fisher viste videre at uttrykket inneholdende s_1, s_2 og r ved hjelp av transformasjoner, bruk av jacobideterminanter til transformasjonene og omskrivninger til trigonometriske funksjoner kunne skrives som:

$$dF = \frac{(1-\rho^2)^{\frac{1}{2}(n-1)}}{\pi\Gamma(n-2)}(1-r^2)^{\frac{1}{2}(n-4)}\frac{d^{n-2}}{d(r\rho)^{n-2}}\big[\frac{arccos(-\rho r)}{\sqrt{1-\rho^2r^2}}\big]dr \qquad (23)$$

jfr. [2, s. 385-387] jfr. [9], som altså gir målefordelingen til r når X og Y kommer fra en bivariat normalfordeling. Et alternativt bevis for målefordelingen til Pearson's r finner du i [12, p. xxxviii]. Vi kan nå lettere studere målefordelingen til Pearson's r under normalfordelingsantagelsen på de underliggende variablene X og Y.
Momentene til fordelingen i likning 23 kan, jfr. [2, s. 390] uttrykkes ved funksjonene:

$$E(r) = \mu_1^{'}(r) = \rho(1 - \frac{1-\rho^2}{2n} + O(n^{-2})) \qquad (24)$$

$$E(r-E(r))^2 = Var(r) = \mu_2(r) = \frac{(1-\rho^2)^2}{n-1}(1+\frac{11\rho^2}{2n}) + O(n^{-3}) \qquad (25)$$

$$Skjevheten = \frac{E(r-E(r))^3}{(\sqrt{\mu_2(r)})^3} = \frac{\mu_3(r)}{(\sqrt{\mu_2(r)})^3} = \gamma_1 = \frac{-6\rho}{n^{\frac{1}{2}}} + o(n^{\frac{-1}{2}}) \qquad (26)$$

$$Kurtosis = \frac{E(r-E(r))^4}{(\sqrt{\mu_2(r)})^4} = \frac{\mu_4(r)}{(\sqrt{\mu_2(r)})^4} = \gamma_2 = \frac{6(12\rho^2-1)}{n} + o(n^{-1}). \qquad (27)$$

Bemerker at likning 24 viser at r er en skjev estimator for ρ, men ved å løse likningen $r = E(r) = \rho - \frac{\rho(1-\rho^2)}{2n}$ m.h.p. ρ oppdager vi en approksimativ forventningsrett estimator $\hat{\rho} = r_{adj} = r(1 - \frac{r^2-1}{2n})$ for ρ. Estimatoren er sub-optimal da den ikke innehar minimumvarians egenskapene. Estimatoren med unikt minst varians(MVU) er gitt ved $r_{MVU} = r *_2 F_1(\frac{1}{2}, \frac{1}{2}; \frac{n-1}{2}; 1-r^2)$, der r,n er definert som tidligere og ${}_2F_1$(a,b;c;z) er den gaussiske hypergeometriske funksjonen jfr. [54, s. 201-211].
Vi konstaterer at skjevheten øker med $|\rho|$ og synker bare med $n^{\frac{-1}{2}}$ og m.a.o. konvergerer fordelingen relativt sakte mot normalfordelingen. Se [52] for mer om forventningsrette estimatorer for Pearson's r.

Egenskaper ved målefordelingen til Pearson's r Det er enkelte karakteristikker, som deles av **alle** målefordelingene til Pearson's r dersom den underliggende fordelingen til (X,Y) er bivariat normal jfr. [5] og momentene ovenfor.
1) Målefordelingen til Pearson's r er ikke tilnærmet normal for små målestørrelser. Når n øker vil fordelingen tilnærmes normalfordelingen saktere om $\rho \neq 0$ enn når $\rho = 0$. Når $\rho \neq 0$ og $|\rho|$ øker blir målefordelingen til Pearsons r for små målinger skjev jfr. likning 26 på forrige side. Spesielt vil målefordelingen til Pearson's r bli negativ skjev(halen skjev mot venstre) når $\rho > 0$. Jo større ρ er jo mer negativ skjev blir målefordelingen til r.
Motsatt blir målefordelingen skjev mot høyre når $\rho < 0$.
2) Når ρ=0 eller ρ=1 er gjennomsnittet til målefordelingen til Pearson's r lik henholdsvis 0 eller 1. E(r) er den forventede verdien til alle mulige verdier av Pearson's r i en målefordeling, som kan bli sett på som gjennomsnittet av alle mulige verdier av Pearson's r. Siden $E(r) = \rho$ når $\rho = 0$ eller $\rho = 1$, jfr. likning 24 på forrige side, anses Pearson's r i disse tilfeller å være en forventningsrett estimator for populasjonskorrelasjonen. Pearson's r er maximumlikelihood(ML)-estimatoren for ρ og dermed også effisient(trenger færre målinger enn noen annen estimator for den samme nøyaktighet), som innebærer at Pearson's r sammenlignet med andre observatorer vil være den beste estimator(minst varians) for populasjonskorrelasjonen ρ når $\rho = 0$ eller $\rho = 1$. Når $\rho \neq 0$ eller $\rho \neq 1$ vil Pearson's r være en ikke forventningsrett estimator for ρ, og skjevheten vil som vi har sett variere med verdien på ρ.
3) Standardfeilen til målefordelingen til Pearson's r basert på store målinger er tilnærmet gitt ved $(\frac{(1-\rho^2)^2}{n})^{\frac{1}{2}}$ jfr. likning 25 på forrige side.

Egenskaper ved store målinger - n høy Dersom den underliggende fordelingen til (X,Y) er bivariat normal vil, som nevnt ovenfor, Pearson's r være maximumlikelihood(ML) - estimatoren for ρ og r er dermed asymptotisk forventningsrett og asymptotisk effisient for alle verdier av ρ. ML-estimatet r er, blant alle tenkelige estimatorer for ρ, den beste og mest eksakte estimatoren for ρ så lenge data er bivariat normalfordelt og målestørrelsen er moderat eller stor. Hva som er moderat eller stor målestørrelse knyttet til Pearson's r når X,Y kommer fra en bivariat normalfordeling avhenger av verdien på populasjonskorrelasjonskoeffisienten ρ. Ved en målestørrelse på n=400 vil fordelingskurvene til r for $\rho = 0.0$ til 0.6 sakte tendere mot normalitet mens for n=400 og $\rho > 0.6$ er det store avvik fra normalfordelingen jfr. [12, p. 33].
For ikke-normale bivariate populasjoner forblir Pearson's r forventningsrett for ρ, men ikke nødvendigvis effisient m.a.o finnes det mer nøyakige/bedre

estimater for ρ enn Pearson's r - Pearson's r vil f.eks. i disse tilfeller ikke nødvendigvis være ML-estimatet for ρ.
Vi har til nå konstatert at Pearson's r er forventningsrett for $\rho = 0$ og $\rho = 1$ for en bivariat normalfordeling og asymptotisk forventningsrett ellers. Pearson's r er videre, uansett fordeling, en konsistent estimator for ρ så lenge målegjennomsnittet, målevariansen og målekovariansen er konsistente og dette er garantert når vi kan bruke loven om store tall. Loven om store tall tilsier at forventningsverdiene til målegjennomsnittet, målevariansen og målekovariansen, når n blir uendelig stor, er lik de tilsvarende populasjonsestimatene og at variansen til målegjennomsnittet, målevariansen og målekovariansen går mot 0 når n blir uendelig stor(en estimator kan altså være (asymptotisk) forventningsrett uten å være konsistent, men vil da uansett få stabil varians når n blir stor nok).

Testing ved bruk av Student's t-fordelingen Dersom vi har en ukorrelert($\rho = 0$) bivariat normalfordeling ser vi av uttykkene 20 på side 21 og 23 på side 22 over at målefordelingen til Pearson's r blir:

$$dF = \frac{1}{B(\frac{1}{2}, \frac{1}{2}(n-2))}(1-r^2)^{\frac{1}{2}(n-4)}dr = \frac{\Gamma(\frac{n-1}{2})}{\sqrt{\pi}\Gamma(\frac{n-2}{2})}(1-r^2)^{\frac{n-4}{2}}dr \tag{28}$$

Denne formen hadde allerede W.S Gosset('Student') antatt eksisterte og beskrevet i [11]. Fordelingsfunksjonen kan finnes ved å studere tabeller over ukomplette Beta-funksjoner, eller mer direkte ved å putte:

$$t = (\frac{(n-2)r^2}{1-r^2})^{\frac{1}{2}} = r\sqrt{\frac{n-2}{1-r^2}} \tag{29}$$

,som dermed reduserer likning 20 på side 21 og likning 23 på side 22 til en Student's t-fordeling med n-2 frihetsgrader jfr. [2, s. 387].
For å bestemme de kritiske verdiene til r under et gitt signifikansnivå trenger vi inversen av transformasjonen for t:

$$r = \frac{t}{\sqrt{n-2+t^2}} \tag{30}$$

Dette resultatet holder også når data ikke er normale, men målestørrelsen er passe høy - se [3, s. 492-493] og 3.1.2 på side 42 om permutasjonstester.
Vi beregner et 100(1-α)% konfidensintervall(KI) for ρ:

$$r \pm t_{\alpha/2} * \sqrt{\frac{1-r^2}{n-2}} \tag{31}$$

, som vi f.eks. kan bruke til å teste hypotesen H_0: $\rho = \rho_0 <=> \rho_0 \in KI$.

Testing ved bruk av målefordelingen til r Dersom $\rho \neq 0$ kan vi fortsatt benytte uttrykket 23 på side 22 ved testing av r. Hypotesetesting og beregning av konfidensintervaller er dog ikke lenger helt trivielt. Hotelling viste at frekvensfunksjonen i likning 23 på side 22 kunne uttrykkes ved en hypergeometrisk funksjon jfr. [2, p. 388][10]. Den hypergeometriske formen Hotelling fant frem til var:

$$dF = \frac{(n-2)dr}{(n-1)\sqrt{2}B(\frac{1}{2}, n-\frac{1}{2})}(1-\rho^2)^{\frac{1}{2}(n-1)}(1-r^2)^{\frac{1}{2}(n-4)}(1-\rho r)^{\frac{3}{2}-n}F(\frac{1}{2}, \frac{1}{2}, n-\frac{1}{2}, \frac{1}{2}(1+\rho r)) \tag{32}$$

Fordelingen konvergerer raskt mot uttrykket 23 på side 22 for små n, og for store n er første termen ofte nok. Feilen ved å stoppe for en gitt n er aldri større enn $\frac{2}{(1-\rho r}$) ganget med den siste termen, som er benyttet.
Ved å integrere uttrykket over term for term får vi fordelingsfunksjonen til Pearson's r og for $r > \rho$ oppdaget Hotelling, jfr. [2, s. 389] jfr. [10], resultatet:

$$1-F(r) = \frac{n-2}{(n-1)\sqrt{2}B(\frac{1}{2}, n-\frac{1}{2})}(M_0+\frac{2M_0-M_1}{4(2n-1)}+\frac{9(4M_0-4M_1+M_2)}{32(2n-1)(2n+1)}+...), \tag{33}$$

der

$$M_k = \int_r^1 (1-\rho^2)^{\frac{1}{2}(n-1)}(1-x^2)^{\frac{1}{2}(n-4)}(1-\rho x)^{\frac{3}{2}+k-n}dx \tag{34}$$

Den asymptotiske formelen for r>ρ blir:

$$1-F(r) \sim \frac{n-2}{(n-1)^2\sqrt{2}B(\frac{1}{2}, n-\frac{1}{2})}\frac{(1-\rho^2)^{\frac{1}{2}(n-1)}(1-r^2)^{\frac{1}{2}(n-2)}}{(r-\rho)(1-\rho r)^{n-\frac{5}{2}}}(1+O(n^{-1})) \tag{35}$$

Ordinatene M_k(k=0,1,2,..) og dermed fordelingsfunksjonen til r kan ikke uttrykkes via simple matematiske funksjoner, men ved hjelp av uttrykk for ordinatene og hovedsaklig bruk av Gregorys numeriske kvadatur formel ved integrasjon jfr. [12, p. 9], laget F.N David tabeller for ordinatene og sannsynlighetsintegralet korrekt opptil 5 desimaler(enkelte unntak for det femte desimalet - se [12]). F.N.David har samlet tabellene for verdiene til n=3,4,...,25, 50, 100, 200, 400 for ρ=0,0.1,0.2,...0.9,1 og r=-1,0.95,0.90,...0,0.05, 0.1,...,1 i slutten av [12]. Ved hjelp av disse tabellene og eventuelt interpolasjon jfr. [12, p. 11-16] kan man utføre hypotesetesting og lage konfidensintervaller. Konfidensintervaller relatert til signifikansnivå på tosidige tester lik 0.05, 0.025, 0.01, 0.005 og ensidige tester på 0.1, 0.05, 0.02, 0.01, kan man, for en gitt $H_0 = \rho$ for n= 3,4,5,6,7,8,9,10,12,15,20,25,50,100,200,400, lese ut av charts 1,2,3 og 4 bakerst i [12].
Konfidensgrenser for en gitt ρ=0.00,0.02,0.04,...,0.96,0.98 for n = 1,2,3,...59,60,62,64,...,

78,80,85,90,95,100,110,...., 190,200,225,250,..,275,300,350,400,..,550,600,700,800,900,1000 for signifikansnivå på tosidige tester lik 0.01,0.02,0.05,0.1,0.2,0.5 og ensidige tester lik 0.25,0.1, 0.05, 0.025, 0.01, 0.005 er gitt i [48] med forklaring til tabellene på sidene 134-136.
I dag benytter vi f.eks. R til å beregne p-verdier og konfidensintervaller relatert til fordelingen, som fremkommer av likning 33 på forrige side.

Fishertransformasjonen Fisher fant frem til en transformasjon av r, som tenderer mot normalitet langt raskere enn r og med varians nærmest uavhengihg av ρ jfr. [13]. Transformasjonen er definert som:

$$F(r) = \frac{1}{2}ln(\frac{1+r}{1-r}) = arctanh(r) \Leftrightarrow r = tanh(F(r)) \tag{36}$$

Setter vi videre:

$$F(\rho) = \frac{1}{2}ln(\frac{1+\rho}{1-\rho}) = arctanh(\rho) \Leftrightarrow \rho = tanh(F(\rho)) \tag{37}$$

kan vi utvikle frekvensfunksjonen til r ved hjelp av polynomer til F(r)-F(ρ)=x og inverse polynomer av n. Fisher ga følgende uttrykk for tettheten f, som det opprinnelig var en slurvefeil i($\frac{2+\rho}{8(n-1)}$ istedenfor $\frac{2+\rho^2}{8(n-1)}$), men som er rettet opp(jfr. [8]) her:

$$f = \frac{n-2}{\sqrt{2\pi(n-1)}}e^{-\frac{1}{2}(n-1)x^2}(1+\frac{1}{2}\rho x+(\frac{2+\rho^2}{8(n-1)}+\frac{4-\rho^2}{8}x^2+\frac{n-1}{12}x^4)+ \tag{38}$$

$$\rho x(\frac{4-\rho^2}{16(n-1)} + \frac{4+3\rho^2}{48}x^2 + \frac{n-1}{24}x^4)+ \tag{39}$$

$$(\frac{4+12\rho^2+9\rho^4}{128(n-1)^2}+\frac{8-2\rho^2+3\rho^4}{64(n-1)}x^2+\frac{8+4\rho^2-5\rho^4}{128}x^4+\frac{28-15\rho^2}{1440}x^6(n-1))+ \tag{40}$$

$$\frac{(n-1)^2}{288}x^8 + ...) \tag{41}$$

Vi kan så finne momentene f.eks om 0 for så å transformere til om forventningen og får da: (merk dette er ikke slik det sto i Fishers papirer og ikke slik det er referert i [2, s. 391], men det var et par feil, som er blitt rettet av Gayen jfr. [8, s. 236]).

$$E(F(r)) = \mu_1^{'}(F(r)) = F(\rho) + \frac{\rho}{2(n-1)}(1 + \frac{5+\rho^2}{4(n-1)} + \frac{11+2\rho^2+3\rho^4}{8(n-1)^2} + ...) \tag{42}$$

$$E(F(r) - F(\rho)) = \mu_1(F(r)) = \frac{\rho}{2(n-1)}(1 + \frac{5+\rho^2}{4(n-1)} + \frac{11+2\rho^2+3\rho^4}{8(n-1)^2} + ...) \tag{43}$$

$$E(F(r) - E(F(r)))^2 = Var(F(r)) = \mu_2(F(r)) = \frac{1}{(n-1)}(1 + \frac{4-\rho^2}{2(n-1)} + \frac{22-6\rho^2-3\rho^4}{6(n-1)^2} + ...) \tag{44}$$

$$Skjevheten = \gamma_1 = \frac{\rho^3}{(n-1)^{\frac{3}{2}}} + ... \tag{45}$$

$$Kurtosis = \gamma_2 = \frac{2}{(n-1)} + \frac{4+2\rho^2-3\rho^4}{(n-1)^2} + ... \tag{46}$$

Vi konstaterer at Fisher-transformasjonen er en tilnærmet variansstabiliserende transformasjon for r på den måten at F(r)-F(ρ) mer eller mindre er uavhengig av ρ, når X og Y følger en bivariat normalfordeling. Dette betyr at variansen til F(r) er tilnærmet konstant for alle verdier av populasjonskorrelasjons-koeffisienten ρ samtidig som vi ser at skjevheten avtar med såpass høy hastighet som $n^{\frac{3}{2}}$(og i utgangspunktet er lav) slik at vi kan si at F(r) - F(ρ) er tilnærmet normalfordelt med forventning og varians gitt av momentene ovenfor. Uten Fisher-transformasjonen vil variansen til r bli mindre når $|\rho|$ nærmer seg 1. Fishertransformasjonen er tilnærmet identitetsfunksjonen når $|r| < 0.3$, men bruk av Fisherapproksimasjonen blir viktigere og viktigere når $|\rho|$ øker over nivået på 0.3.
Vi ser videre av variansen og skjevheten at en noe røffere, men enda enklere approksimering er å sette $E(F(r) - F(\rho)) = \frac{\rho}{2(n-1)}$ og $Var(F(r) - F(\rho)) = \frac{1}{n-1} + \frac{4-\rho^2}{2(n-1)^2}$, som for lave ρ er tilnærmet lik $\frac{1}{n-1} + \frac{2}{(n-1)^2} = \frac{1}{n-3}$. Approksimasjonene er gode når n>50 mens de 'fulle' uttrykkene er funnet å fungere tilfredsstillende helt ned til n=11 ved bruk av asymptotisk formel - se f.eks [12].
Det er gjort ytterligere dog mindre kjente forbedringer av Fishertransformasjoen av Hotelling [10] ($F(r)^* = F(r) - \frac{3F(r)+r}{4n}$) og så av Ruben(1966). Transformasjoner, som er laget for å stabilisere variansen(f.eks gjøre den uavhengig av en populasjonsparameter(ρ i Fishertransformasjonen), vil også typisk være med å normalisere fordelingen transformasjonen blir brukt på - Fisher z-transformasjonen er et eksempel på dette. Uansett for å finne en best mulig transformasjon av denne type må vi kjenne den underliggende fordelingen.

Testing ved hjelp av Fishertransformasjonen

- H_0: $\rho = \rho_1$
  Når vi ikke skal teste H_0: $\rho = 0$, men H_0: $\rho = \rho_1$, der $\rho_1 \neq 0$, eller lage konfidensintervaller for ρ basert på observatoren vi har beregnet for r, er det relativt tungvint og benytte den eksakte målefordelingen da den for mange er noe krevende og interpretere. Hvis (X,Y) kommer fra en bivariat normalfordeling og hvis (X_i, Y_i) - parene for i=1,2,..n, som brukes til å beregne r, er uavhengige så har vi sett at F(r), for et visst nivå på n igjen avhengig av nivået på ρ, er tilnærmet normalfordelt med forventning og standardfeil:
  E[F(r)] = F(ρ) = $\frac{1}{2} ln \frac{1+\rho}{1-\rho}$
  SE = $\frac{1}{\sqrt{n-3}}$
  , der n er målestørrelsen. Det er ikke noe i veien for å bruke de fulle uttrykkene for E(F(r)) og SE slik at vi kan utføre hypotesetesting og konstruere konfidensintervaller for ρ for n helt ned til 11.
  Tester vi hypotesen H_0: $\rho = \rho_1$, der $\rho_1 \neq 0$ vil vi, som følge av ovenfornevnte, kunne bruke formelen:

$$z = \frac{F(r) - F(\rho_1)}{SE} \tag{47}$$

  , der F(r) er den transformerte verdien til Pearson's r fra et utvalg og $F(\rho_1)$ er den transformerte verdien til ρ_1. z-observatoren beskriver hvor mange standardfeil F(r) er unna $F(\rho_1)$ gitt at nullhypotesen er sann. P-verdien blir sjansen for å observere en |z| større enn den |z| vi har observert. Basert på en tosidig test med signifikansnivå 0.05 er den kritiske verdien tilnærmet lik 1.96 og m.a.o. forkaster vi H_0 dersom |z|>1.96.

Konfidensintervaller Basert på en observator for r fra et utvalg på n av populasjonen kan vi lage vi et 95% konfidensintervall for populasjonskorrelasjonskoeffisienten ρ. Transformasjonen sammen med dens inverse
r = $\frac{exp(2F(r))-1}{exp(2F(r))+1} = tanh(F(r))$
kan bli brukt til å konstruere konfidensintervaller. Vi har at
$z = (F(r) - F(\rho))\sqrt{n-3} \sim$ N(0,1)
For å finne et konfidensintervall for ρ beregner vi først et 100(1-α)% konfidensintervall for F(ρ):

$$F(r) \pm z_{\alpha/2} SE \tag{48}$$

Den inverse Fishertransformasjonen får intervallet tilbake til korrelasjonsskala og et 100(1-α)% konfidensintervall for ρ blir:

$$[tanh(F(r) - z_{\alpha/2}SE), tanh(F(r) + z_{\alpha/2}SE)] \tag{49}$$

2.6 Robusthet

En statistisk prosedyre, som ikke er sensitiv mot avvik fra antagelsene den statistiske prosedyren hviler på kalles robust - et uttrykk, som ble introdusert av G.E.P.Box [15]. Studier knyttet til robusthet har hatt mye fokus på variansanalyse - se f.eks [3]. I likhet med mange brukte observatorer er observatoren r ikke robust. Pearson's r er ikke robust mot ulike fordelinger av variablene X og Y. Eksakte tester og asymptotiske tester basert på en Fisher-transformasjon av data kan brukes hvis data er tilnærmet normalfordelte, men kan være misledende ellers.
A.K. Gayen har sett på både robustheten til målekorrelasjonskoeffisienten r og Fishertransformasjonen av r når x,y avviker fra en bivariat normalfordeling jfr. [8]. Når populasjonskorrelasjonskoeffisienten $\rho = 0$, og spesielt når variablene er uavhengige, er fordelingen til r robust for målestørrelser så lavt som n=11, men for store verdier av ρ er avvik fra normalteori et problem.
Pearson's r er heller ikke resistent mot uteliggere og spesielt er dette tilfelle ved et lite utvalg hvor en uteligger fort både kan påvirke størrelse, men også retning til Pearson's r.

2.7 To eller flere uavhengige utvalg

2.7.1 To uavhengige utvalg - H_0: $\rho_1 = \rho_2$

Ikke sjelden støter man i praksis og forskning på spørsmålet om hvorvidt en korrelasjon i utvalg 1(n_1) er ulik en korrelasjon man fikk i utvalg 2(n_2). Til sammenligning med nullhypotesen $\rho = \rho_1$ har vi nå å gjøre med en statistisk inferenstest for forskjellen mellom to korrelasjoner beregnet fra to uavhengige målinger/utvalg 1 og 2. Ved testing av $\rho = \rho_1$ har vi å gjøre med en statistisk inferenstest for forskjellen mellom en korrelasjon beregnet fra et enkelt utvalg og en spesifikk 'standard' korrelasjon valgt av forskerne. Dersom vi har målinger fra to populasjoner med korrelasjonene ρ_1 og ρ_2 vil en metode å teste hypotesen H_0: $\rho_1 = \rho_2$ typisk være å studere forskjellen mellom målekorrelasjonene delt på standardfeilen til denne

forskjellen og knytte dette forholdet opp mot en normalfordeling jfr. [12, p. xxii]. Det beskrevne forholdet blir altså(se likning 25 på side 22 for standardavvik):

$$\frac{r_1 - r_2}{\sqrt{\frac{(1-r_1^2)^2}{n_1-1} + \frac{(1-r_2^2)^2}{n_2-1}}} \tag{50}$$

Prosedyren er adekvat ved n meget høy forutsatt at den hypotetisk felles $|\rho|$ ikke er for nær 1 og kan være svært unøyaktig ellers. Målefordelingen til forholdet over er ukjent og formen vil være meget vanskelig å oppdage da r_1 og r_2 følger ulike ikke-normale fordelinger avhengig av verdiene på n_1 og n_2 og av den ukjente felles parameteren ρ.
R.A.Fisher's F(r)-transformasjon er som vi vet nær normalfordelt og der standardfeilen er nær uavhengig av ρ. Hvis vi har utvalg fra flere populasjoner kan vi utføre hypotesetesting relatert til populasjonenes ρ's ved å bruke tester fra normalteori på F(r)'ene. For å undersøke om korrelasjonen i utvalg 1 er forskjellig fra den i utvalg 2 kan vi utføre en z-test i henhold til:

$$z = \frac{F(r_1) - F(r_2)}{\sqrt{SE^2_{F(r_1)} + SE^2_{F(r_2)}}} \tag{51}$$

En alternativ test foreslått av E.S.Pearson, som benytter målefordelingen til r, er beskrevet og bevist i [12, p. xxviii-xxxi]. Testen kan benyttes dersom n_1 og n_2 er lave(også lavere enn 11) og ikke er veldig ulike og testen krever ingen transformasjon på variabler, men forutsetter at begge de underligende bivariate fordelingene er normale. Jeg vil illustrere testen ved hjelp av chart 1 fra [12] jfr. figurene 1 og 2. Konfidensbeltene i chart 1 er laget med en konfidensgrad på 90%. Vi benytter de observerte verdiene r_1 og r_2 til å lese av, fra de tilhørende konfidensbeltekurver for n_1 og n_2(i praksis gjør vi dette i R) størrelsene ρ_{a_1}, ρ_{b_1}, ρ_{a_2}, ρ_{b_2}, som vist i figur 2. Størrelsene er definert via den eksakte fordelingen til Pearson's r, jfr. likning 23 på side 22 for gitte verdier av r_1 og r_2. De avleste størrelsene er henholdsvis nedre(ρ_{a_1},ρ_{b_1}) , og øvre(ρ_{a_2},ρ_{b_2}) konfidensintervallgrense for ρ_1 og ρ_2 relatert til de observerte r_1 i utvalg n_1 og r_2 i utvalg n_2.
Vi må skille mellom hypotesene 1) H_0: $\rho_1 = \rho_2$(alternativ hypotese: $\rho_1 > \rho_2$ eller $\rho_1 < \rho_2$), 2) H_0: $\rho_1 \geq \rho_2$(alternativ hypotese: $\rho_1 < \rho_2$) og tilsvarende for H_0: $\rho_1 \leq \rho_2$(alternativ hypotese: $\rho_1 > \rho_2$).
Det er en nødvendig betingelse for å forkaste nullhypotesen at de to konfidensintervallene overlapper hverandre. For versjon 1) forkaster vi

H_0: $\rho_1 = \rho_2$ hvis $\rho_{b_2} > \rho_{a_1}$, eller $\rho_{b_1} > \rho_{a_2}$. Risikoen ved å forkaste H_0 når den er sann er tilnærmet bare 0.02 forutsatt at n_1 og n_2 ikke er svært ulike.
For versjon 2) forkaster vi H_0: $\rho_1 \geq \rho_2$ hvis $\rho_{b_2} > \rho_{a_1}$ og tilsvarende forkaster vi H_0: $\rho_1 \leq \rho_2$ hvis $\rho_{b_1} > \rho_{a_2}$ i versjon 3). Type 1-feilen, sjansen for å forkaste H_0 når den er sann, er nå tilnærmet lik 0.01.
Jeg bemerker at det vil være vågalt og akseptere en hypotese, som blir bevist med en av testene her, når tilgangen på data er liten. Derimot vil ytterligere undersøkelser mulig kunne bekrefte utfallet av testen.

2.7.2 Avhengige korrelasjoner - H_0: $\rho_1 = \rho_2$

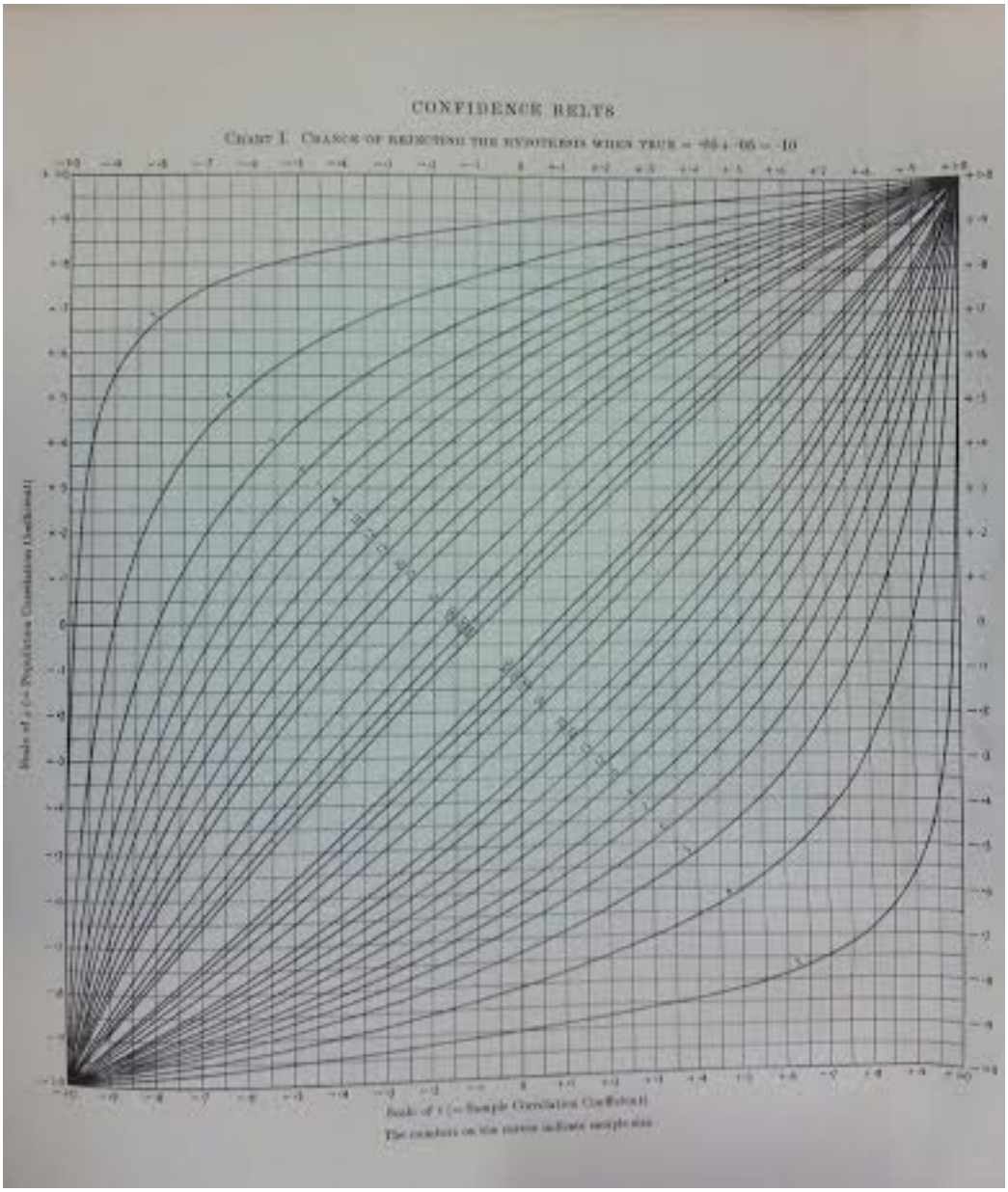
Anta at vi måler 3 variabler(f.eks. X=VO2max, Y=treningsmengde, Z=hemoglobinnivå) på en gruppe skiløpere og gjør tilsvarende neste år(X'=VO2max, Y'=treningsmengde og Z'=hemoglobinnivå).
I tilfelle 1 nedenfor vil jeg benytte de tre mulige korrelasjoner første året (r_{xy}, r_{xz}, r_{yz}) og i tilfelle 2 nedenfor vil jeg benytte de seks mulige korrelasjonenene mellom X, Y, X' og Y' ($r_{xy}, r_{xx'}, r_{xy'}, r_{yx'}, r_{yy'}, r_{x'y'}$) , der man både hva gjelder tilfelle 1 og tilfelle 2 ikke generelt kan hevde at de ulike korrelasjonene er uavhengige av hverandre - vi kan derfor ikke alltid bruke den uavhengige testen i foregående punkt for nullhypotesetesting jfr. Steiger [41].

Tilfelle 1. Er variabelen X relatert til variabel Y på en annen måte enn til variabel Z? Vi ønsker altså å sjekke hvorvidt VO2max er mer relatert til treningsmengde sammenlignet med hemoglobinnivå. Nullhypotesen blir ρ_{xy}=ρ_{xz}. Steiger jfr. [41] viser at William's t-test kan bli brukt med fornuft når målestørrelsen oppfyller kravet n > 20. Formelen for t-observatoren til William med (n-3) frihetsgrader er:

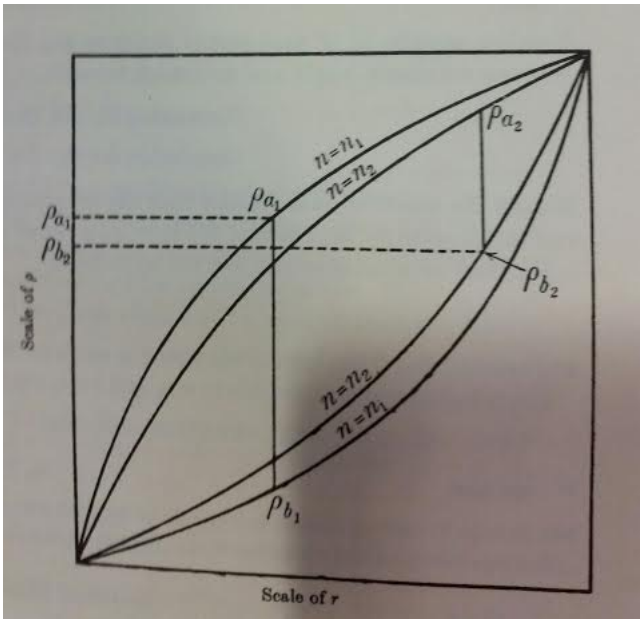
$$t = (r_{xy} - r_{xz})\sqrt{\frac{(n-1)(1+r_{yz})}{2((n-1)/(n-3))|R| + \bar{r}^2(1-r_{yz})^3}} \qquad (52)$$

, der $|R| = 1 - r_{xy}^2 - r_{xz}^2 - r_{yz}^2 + 2r_{xy}r_{xz}r_{yz}$ og $\bar{r} = \frac{r_{xy}+r_{xz}}{2}$

Tilfelle 2. Er relasjonen mellom variablene X og Y ulik relasjonen mellom variablene X' og Y'? Vi ønsker altså å teste om treningsmengde er mer relatert til VO2max første året smln. med andre året. Nullhypotesen blir ρ_{xy}=$\rho_{x'y'}$. Steiger jfr. [41] viser at man kan bruke



Figur 1: Chart1 fra [12]



Figur 2: Utdrag fra Chart 1

Dunn og Clark's z-test hvis n > 20 der observatoren for den tilnærmede standardnormaltfordelte variabelen z er gitt ved:

$$z = (F(r_{xy}) - F(r_{x'y'})) * \sqrt{\frac{n-3}{2-2\bar{s}}}, \quad \bar{s} = \frac{\psi}{(1-\bar{r}^2)^2}, \quad \bar{r} = \frac{r_{xy} + r_{x'y'}}{2} \tag{53}$$

, der $F(r_{xy})$ og $F(r_{x'y'})$ er transformerte Fisher's $F(r)$-verdier av Pearson's r_{xy} og $r_{x'y'}$ og $\psi = 0.5(((r_{xx'} - r_{yx'}\bar{r})(r_{yy'} - r_{yx'}\bar{r})) + ((r_{xy'} - r_{xx'}\bar{r})(r_{yx'} - r_{xx'}\bar{r})) + ((r_{xx'} - r_{xy'}\bar{r})(r_{yy'} - r_{xy'}\bar{r})) + ((r_{xy'} - r_{yy'}\bar{r})(r_{yx'} - r_{yy'}\bar{r})))$.

2.8 Spesialtilfeller av Pearson's r

Jeg skal her, jfr. [5], gi en rask oversikt over andre parametriske korrelasjonsindekser, som alle bare er spesialtilfeller av Pearson's r. Siden alle korrelasjonene er spesialtilfeller av Pearson's r jfr. [35] 2.utgave er alle nullhypotesene, brukt ved testing av Pearson's r, generelt også anvendelige her. Jeg bemerker at alle beregninger gjort i denne delseksjonen i all hovedsak er matematiske forenklede versjoner av formelen for Pearson's r. Bemerker også at absoluttverdien til disse korrelasjonene tenderer mot å være mindre enn 1, fordi formen på fordelingene til X og Y ofte ikke er symmetriske, eller har samme

form jfr. diskusjonen under 2.3 på side 17.

2.8.1 Den biserielle punktkorrelasjonen mellom en binær og kontinuerlig variabel - r_{pb}

Vi bruker formelen for Pearson's r rett frem:

$$r = r_{pb} = \frac{n\sum_i X_i Y_i - \sum_i X_i \sum_i Y_i}{\sqrt{n\sum_i X_i^2 - (\sum_i X_i)^2}\sqrt{n\sum_i Y_i^2 - (\sum_i Y_i)^2}} \tag{54}$$

$$\Updownarrow$$

$$r = r_{pb} = \frac{\sum_i X_i Y_i - \frac{\sum_i X_i \sum_i Y_i}{n}}{\sqrt{\sum_i X_i^2 - \frac{(\sum_i X_i)^2}{n}}\sqrt{\sum_i Y_i^2 - \frac{(\sum_i Y_i)^2}{n}}} \tag{55}$$

Lar X være den binære variabelen refererende til mann eller kvinne kodet som henholdsvis 0 og 1. X_i kan da ta verdiene 0 og 1 for individene i=1,2,...,n. Lar Y_i være den kontinuerlige variabelen, som f.eks. refererer til maks VO2-opptak for de samme individene i=1,2,...,n. Bemerker at det ikke spiller noen rolle hvordan vi numerisk koder mann og kvinne bare de numeriske verdiene er ulike, men dersom vi skifter rekkefølgen på kodingen(f.eks. mann=1 og kvinne=0) vil fortegnet på korrelasjonen skifte fortegn.
Den binære variabelen X har ikke den samme marginalfordelingen som Y og maksimumsverdien til $|r_{pb}|$ vil derfor alltid være mindre enn 1.
Inferens på r_{pb} blir helt lik, som for r.
Jeg bemerker at inferens på r_{pb} essensielt gir samme informasjon som den kjente t-testen for to uavhengige utvalg basert på:

$$t = \frac{\bar{Y_1} - \bar{Y_2}}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \tag{56}$$

og

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} \tag{57}$$

, der antall frihetsgrader er $(n_1 + n_2 - 2)$, variansen $s_k^2 = \frac{\sum_i (Y_i - \bar{Y_k})^2}{n_k - 1}$ for k=1,2 med referanse til de to populasjonen menn og kvinner og der $\bar{Y_1}$ og $\bar{Y_2}$ er gjennomsnittene i de to populasjonene. Summeringen går over henholdsvis n_1 menn og n_2 kvinner.

Relasjonen mellom r_{pb} og den observerte t-observatoren observert ved t-testen for to uavhengige utvalg kan beskrives ved:

$$t = \frac{r_{pb} * \sqrt{n_1 + n_2 - 2}}{\sqrt{1 - r_{pb}^2}} \Leftrightarrow r_{pb}^2 = \frac{t^2}{(n_1 + n_2 - 2) + t^2} \tag{58}$$

Dersom du erstatter r_{pb}, i uttrykket for t rett ovenfor, med uttrykket for Pearson's r, vil du ved manipulasjon få det samme uttrykket som brukes for t i t-testen for to uavhengige utvalg. Forskjellen på testene ligger i fokuset på det som testes. Korrelasjonstesten er opptatt av om det foreligger en korrelasjon mens t-testen fokuserer på om det er en forskjell på de to gjennomsnittene - dette er essensielt altså det samme.

2.8.2 ϕ-koeffisienten for korrelasjon mellom to binære variable

ϕ-koeffisienten er akkurat som r_{pb} et annet spesialtilfelle av Pearson's r og vi bruker formelen for Pearson's r rett frem akkurat som for r_{pb}, men slik at begge variablene (X og Y) er binære og hver seg altså bare kan ta to verdier. F.eks undersøker vi sammenhengen mellom hvorvidt man blir tatt opp på en skole i forhold til om man er mann eller kvinne.
Selvom den teoretiske rekkevidden til Pearson's r varierer fra 1 til -1 vil rekkevidden til ϕ i absoluttverdi, som oftest være mindre enn 1 med mindre $A_X = A_Y = 0.5$, der A_X og A_Y er andelen av observasjonene, som får verdien 1 på henholdsvis variabel X og variabel Y. Det er mulig for ϕ og være 1, men ikke -1 når $A_X = A_Y \neq 0.5$. Motsatt kan ϕ være -1 og ikke 1 når $A_X = B_Y \neq 0.5$, der B_X og B_Y er andelen av observasjoner som scorer 0 på henholdsvis variabel X og variabel Y. Caroll jfr. [38] viste, under forutsetning av at $A_X A_Y B_X B_Y \neq 0$, at maksimumsverdien til ϕ er gitt ved $\sqrt{\frac{A_X B_Y}{A_Y B_X}}$ når $A_X < A_Y$ og $\sqrt{\frac{A_Y B_X}{A_X B_Y}}$ når $A_Y < A_X$. Vi ser at jo større forskjellen er på A_X og A_Y jo mindre blir maksimumsverdien til ϕ.
Se nærmere om ϕ-koeffisienten under jfr. 4.3.4 på side 62.

2.8.3 Rådata kontra transformerte data

Dersom man beregner Pearson's r basert på to kontinuerlige variable får man ikke den samme verdien på r_{pb} eller ϕ dersom man konverterer en eller begge de kontinuerlige variablene til en eller to binære variable. Generelt tenderer Pearson's r mot å være høyere enn r_{pb} og ϕ i disse tilfellene. Fallet i verdien til r_{pb} og ϕ skyldes hovedsakelig tap av informasjon og presisjon når en eller to kontinuerlige variable blir dikotomisert. M.a.o. vil typen av assosiasjon

mellom to variable gitt ved r_{pb} og ϕ kunne være forskjellig avhengig av om r_{pb} og ϕ er beregnet på rådata eller konverterte data. Hvis r_{pb} og ϕ er beregnet på rådata(sanne binære kategorier) inneholder de den lineære relasjonen mellom variablene på lik linje med Pearson's r, men dersom r_{pb} og ϕ beregnes etter at data er konvertert til henholdsvis 1 og 2 binære variable blir de mål på en monotonisk relasjon, som beskriver i hvilken grad variablene beveger seg i samme eller motsatt retning. En positiv/negativ monotonisk relasjon forteller at begge variablene beveger seg i samme/motsatt retning i et ikke-lineært mønster.
Det skjer at forskere ønsker å konvertere rådata fra kontinuerlige variable. Det kan eksistere uteliggere, som påvirker relasjonen mellom variablene. Denne påvirkningen kan fjernes ved f.eks. å rangere data slik at variabelen blir ordinal, eller i 'verste fall' dikotomisere den. Andre ganger kan det være at enkelte data ikke er blitt oppgitt med spesifikke verdier(f.eks vekt>68) og igjen kan det være aktuelt å dikotomisere eller polytomisere variabelen.

Den biserielle korrelasjonskoeffisienten mellom en binær og en kontinuerlig variabel - r_{bis} - et estimat for r_{pb}

I dette tilfelle har vi en kontinuerlig variabel Y og en variabel X, som er dikotomisert ved at den kun er målt på to verdier, men der variabelen i utgangspunktet er kontinuerlig og normalfordelt. r_{bis} er en korrelasjonsindeks, som finner korrelasjonen mellom den dikotomiserte variabelen X(typisk utvalg1 og utvalg 2) og den kontinuerlige variabelen Y. r_{bis} finnes ved hjelp av formelen:

$$r_{bis} = \frac{\bar{Y}_2 - \bar{Y}_1}{s_Y} * \frac{n_2 n_1}{\lambda(n_1 + n_2)^2} \tag{59}$$

, der $\bar{Y}_1$ og $\bar{Y}_2$ er gjennomsnittene til Y for utvalg 1 og utvalg 2 med tilhørende målestørrelser n_1 og n_2, s_Y er det samlede standardavviket for Y og der λ er ordinatet(andrekoordinaten) for den standardiserte normalfordelingen med tetthet ϕ relatert til punktet $\frac{n_2}{n_1+n_2}$. Dersom det er like stor andel i hver gruppe så blir ordinatet $\lambda = \phi(0) = \frac{1}{\sqrt{(2\pi)}} = 0.398942$.
Generelt er:

$$r_{bis} = r_{pb}\sqrt{\frac{n_2 n_1(n_1 + n_2 - 1)}{\lambda^2(n_1 + n_2)^2}} \tag{60}$$

og siden $\sqrt{\frac{n_2 n_1(n_1+n_2-1)}{\lambda^2(n_1+n_2)^2}} \geq 1.25$ vil r_{bis} alltid være større enn r_{pb} jfr. Glass og Hopkins [39] og dette forklarer hvorfor relasjonen mellom to kontinuerlige variable, som kommer fra en bivariat normalfordeling, blir mindre hvis en av dem dikotomiseres.

Tabell 3: Kontingenstabell for to binære variable 1

	Binær variabel 1 - X	
Binær variabel 2 - Y	Ja	Nei
Ja	a	b
Nei	c	d

Tetrachoric korrelasjonskoeffisient mellom to binære variable - r_{tet} - et estimat for ϕ

Brukes når to variable (X og Y) hver seg er målt på to verdier selvom den underliggende fordelingen til de to variablene er bivariat normal. Relasjonen r_{tet} mellom disse to dikotomiserte variablene er gitt ved:

$$r_{tet} = \frac{ad - bc}{\lambda_X \lambda_Y n^2} \tag{61}$$

, der a,b,c,d jfr. tabell 3 viser til antall observasjoner som faller i hver av de 4 mulige kategoriene, som de to dikotomiserte variablene gir opphav til. λ_X og λ_Y er ordinatene til henholdsvis variabel X og Y. Sammenligner vi r_{tet} og ϕ kalkulert på de samme data vil vi se at ϕ er lavere. Bemerker at r_{tet} ikke vil være noe godt estimat for relasjonen mellom de kontinuerlige variablene X og Y med mindre målefordelingen er over 400 jfr. [39].
Generelt bør man unngå å konvertere kontinuerlige data med mindre det er praktiske , eller særlige grunner til å gjøre dette. Det viser seg at r_{bis} og r_{tet} sjeldent blir brukt i praksis og bør benyttes med forsiktighet da de er hypotetiske korrelasjoner. Den spesielt interesserte leser kan lese mer om nullhypotesetesting relatert til r_{bis} og r_{tet} i [39, p. 368-369].

2.8.4 η-koeffisienten for korrelasjon mellom en multippel og en kontinuerlig variabel - η

I henhold til Wherry [40] er η også et spesialtilfelle av Pearson's r hvis observasjonsverdiene til den multiple variabelen for hvert tilfelle innenfor hver kategori erstattes av sitt gjennomsnitt.
I motsetning til korrelasjonsindeksene vi har diskutert til nå beskriver η-koeffisienten kun styrken av relasjon mellom variable. Retningen(+/-) på styrken er ikke av interesse, fordi kategoriene til den multiple variabelen ikke representerer noen ordnet rekkefølge(f.eks. skimerker). At retningen på styrken ikke er av betydning gjør at vi også kan bruke η-koeffisienten til å måle krumlinjet sammenheng mellom en multippel variabel

Tabell 4: Glidertest - korrelasjonskoeffisienten η

Smln. av glidhastighet(m/s) i en bakke med lite helling					
m/s Glider A	Glider A	m/s Glider B	Glider B	m/s Glider C	Glider C
2.5	2.5	3.2	3.13	3.5	3.33
2.6	2.5	3.4	3.13	3.2	3.33
2.7	2.5	2.6	3.13	3.0	3.33
3.2	2.5	3.2	3.13	3.0	3.33
2.8	2.5	3.9	3.13	3.6	3.33
2.4	2.5	2.7	3.13	2.9	3.33
2.1	2.5	3.1	3.13	3.3	3.33
2.0	2.5	2.9	3.13	4.1	3.33
2.5	2.5	3.4	3.13	3.2	3.33
2.2	2.5	2.9	3.13	3.5	3.33
n_A=10	n_B=10	n_C=10			
$\bar{Y}_A = 2.5$	$\bar{Y}_B = 3.13$	$\bar{Y}_C = 3.33$			
s_A=0.36	s_B=0.38	s_C=0.36			

og en kontinuerlig variabel. η-koeffisienten, som ofte refereres til som en korrelasjonsratio, brukes vanligvis i en variansanalyse-sammenheng. Jeg illustrerer sammenhengen mellom ANOVA(variansanalyse) og η-koeffisienten med et hypotetisk eksempel. Vi tester hastigheten på et par ski tre ganger hver gang med ny glider. Vi har 10 ulike utøvere, som tester hver av gliderne. Resultatene av testen ses i tabell 4 og en variansanalyse på dataene i den nevnte tabellen fremkommer av tabell 5 på neste side. η-koeffisienten er gitt ved :

$$\eta = \sqrt{\frac{SS_B}{SS_{Total}}} \tag{62}$$

I glidertesteksempelet konstaterer vi at $\eta = \sqrt{\frac{3753}{7375}} = 0.71$, som er signifikant forskjellig fra null i henhold til den observerte F-verdien i tabell 5 på neste side(F=13.99 og p<0.05 for tosidig test). I henhold til Wherry [40] er η også et spesialtilfelle av Pearson's r hvis observasjonsverdiene til den multiple variabelen for hvert tilfelle innenfor hver kategori erstattes av sitt gjennomsnitt (de observerte hastighetene relatert til gliderne A,B og C blir erstattet med henholdsvis 2.5, 3.13 og 3.33) jfr. 4. Beregner man så Pearson's r mellom Glider og hastighet på disse dataene får vi $|r| = 0.71 = \eta$.

Tabell 5: Glidertest - anova

	SS - Sum kvadratavvik	Frihetsgrader	$\bar{SS}$	F	Signifikans
SS_B - mellom grupper	3.753	2	1.876	13.987	0.000
SS_W - innenfor gruppene	3.622	27	0.134		
Total	7.375	29			

2.8.5 Korrelasjon mellom en binær og multippel variabel - r_{MD}

Wherry jfr. [40] har redegjort for korrelasjonen r_{MD} mellom en binær og multippel variabel. Vi trenger å omkode den multiple variabelen for å kunne beregne r_{MD} ved hjelp av formelen for Pearson's r. Omkodingen av den multiple variabelen skjer på eksakt samme måte, som beskrevet for den multiple variabelen i seksjon 2.8.4 på side 37 ved beregning av η-koeffisienten. Siden de k kategoriene for den multiple variabelen kan rangeres i en hvilken som helst rekkefølge vil retningen(+/-) på r_{MD} basert på Pearson's r formelen ikke være av interesse slik at vi kun ser på absoluttverdien. Den passende nullhypotesetesten for r_{MD} er observatoren χ^2 = n*r^2_{MD} med k-1 frihetsgrader.

2.8.6 Korrelasjon mellom en binær og en ordinal variabel - r_{DR}

Wherry jfr. [40] har redegjort for korrelasjonen r_{DR} mellom en binær og ordinal variabel. r_{DR} kan beregnes ved direkte bruk av formelen for Pearson's r. Nullhypotesen $\rho = 0$ kan utføres ved hjelp av observatoren z = $r_{DR}(n-1)^{\frac{1}{2}}$ når n>30 da r_{DR} da er tilnærmet normalfordelt med forventing 0(under nullhypotesen) og standardavvik $(n-1)^{\frac{1}{2}}$ og z dermed tilnærmet standardnormalfordelt.

2.8.7 Korrelasjon mellom en multippel og ordinal variabel - r_{MR}

Wherry jfr. [40] har redegjort for korrelasjonen r_{MR} mellom en multippel og ordinal variabel. Vi trenger å omkode den multiple variabelen for å kunne beregne r_{MD} ved hjelp av formelen for Pearson's r. Omkodingen av den multiple variabelen skjer på eksakt samme måte, som beskrevet for den multiple variabelen i seksjon 2.8.4 på side 37 ved beregning av η-koeffisienten. Siden de k kategoriene for den multiple variabelen kan rangeres i en hvilken som helst rekkefølge vil retningen(+/-) på r_{MR} basert på Pearson's r formelen ikke være av interesse slik at vi kun ser på absoluttverdien. Den

passende nullhypotesetesten for r_{MR} er observatoren χ^2 = (n-1)*r^2_{MR} med k-1 frihetsgrader.

2.8.8 Korrelasjon mellom en ordinal og en kontinuerlig variabel - r_{RI}

Wherry jfr. [40] har redegjort for korrelasjonen r_{RI} mellom en multippel og kontinuerlig variabel. r_{RI} kan beregnes ved direkte bruk av formelen for Pearson's r. Nullhypotesen $\rho = 0$ kan utføres ved hjelp av observatoren z = $r_{RI}(n-1)^{\frac{1}{2}}$ når n>30 da r_{RI} da er tilnærmet normalfordelt med forventing 0(under nullhypotesen) og standardavvik $(n-1)^{\frac{1}{2}}$ og z dermed tilnærmet standardnormalfordelt.

3 Permutasjonstester og bootstrapping - Pearson's r

Disse ikke-parametriske tilnærmingene for å utføre inferens, knyttet til korrelasjonskoeffisienten Pearson's r, avhenger ikke av fordelingen på data.
Vi skal se på permutasjonsfordelinger og hvordan disse kan brukes til å teste hypotesen H_0: $\rho = 0$.
Vi skal også kort se hvordan man kan benytte ikke-parametrisk, og parametrisk, bootstrapping til å lage konfidensintervaller for parameteren ρ og herunder utføre hypotesetesting.
Hypotesetesting knyttet opp mot permutasjonsfordelinger tar utgangspunkt i hele fordelingen mens hypotesetesting knyttet opp mot bootstrapping tar utgangspunkt i parameterestimatene.

3.1 Permutasjonstester

Generelt ved bruk av permutasjonstester må man spesielt skille på to modeller [44]:
1) Randomiseringsmodellen. Tilgjengelige subjekter blir tilfeldig delt i to grupper - f.eks behandling/ikke behandling. Prosedyrene brukt under modellen kalles typisk randomiseringstester, eller randomiseringsintervaller.
2) Populasjonsmodellen. Subjekter blir tilfeldig valgt fra to uavhengige populasjoner. Prosedyrene brukt under modellen kalles typisk permutasjonstester, eller permutasjonsintervaller.
Ved disse generelle permutasjonstestene lar det seg også gjøre å utlede konfidensintervaller, men dette krever langt flere beregninger enn bare å utføre

hypotesetesting og det henvises til [44, s. 680 og 682] for utledning og bevis. I oppgaven her skal vi se på permutasjonsfordelinger i en korrelasjonskontekst. Ved bruk av permutasjonstester relatert til en korrelasjonskoeffisient, som Pearson's r, er det forutsatt at det ikke er noen orden eller gruppering på dataparene. Det er m.a.o. kun mulig å teste hypotesen H_0: $\rho = 0$. Vi gjør typisk målinger av to variable X og Y på samme subjekter(f.eks. langrennsløpere) i et tilfeldig utvalg fra en populasjon, for deretter å utføre permutasjonstesten. Det blir ikke aktuelt å lage konfidensintervaller.

3.1.1 Eksakt permutasjonsfordeling

Teorien ble utviklet av R.A. Fisher og E.J.G. Pitman i 1930-årene.
For å benytte en permutasjonstest er det viktig at alle rekkefølger/rangeringer har lik sjanse for å forekomme under nullhypotesen - i såfall får vi eksakte signifikansnivåer. En permutasjonstest er en type statistisk signifikanstest der fordelingen til observatoren, som testes under nullypotesen, oppdages ved å beregne alle mulige verdier av observatoren som skal testes. Vi skal her studere permutasjonstester relatert til Pearson's r, men i kapittel 4 vil vi studere tilsvarende permutasjonstester relatert til Kendall's t og Spearman's r_s.
Vi antar at det eksisterer en populasjon der populasjonskorrelasjonen ρ mellom X og Y er null. Permutasjonsfordelingen oppdages da ved at vi velger ut n par (X,Y) av populasjonen. Vi holder y'ene faste og permuterer x'ene på n! måter. Vi beregner så for hver av de n! kombinasjonene av n par tilhørende Pearson's r. Da vil alle mulige verdier av Pearson's r basert på målestørrelsen n og alle sannsynligheter for å observere hver verdi av Pearson's r oppdages. Vi kan dermed endelig konstruere en målefordeling for Pearson's r, gitt ρ=0, som viser den funksjonelle relasjonen mellom verdiene til Pearson's r vist på x-aksen og den korresponderende sannsynlighet vist på y-aksen.
Bruker vi målefordelingen for Pearson's r slik den er beskrevet her kan vi altså finne korresponderende sannsynlighet hvis vi kjenner en enkelt Pearson's r fått fra en enkelt måling. Mer interessant er det å teste hvorvidt den observerte Pearson's r fått fra en enkelt måling er signifikant forskjellig fra 0. Intuitivt vet vi at det skal være en liten sjanse og oppdage store Pearson's r når ρ=0 og motsatt er det en stor sjanse for å oppdage en liten verdi på Pearson's r når ρ=0. Den ensidige høyresidige p-verdien til testen beregnes som andelen av de n! permuterte beregnede Pearsons'r, som er større eller lik den observerte Pearson's r(alternativt lavere eller lik ved venstresidig test). Det er klart, utfra at permutasjonsfordelingen er diskret, at ovenfornevnte p-verdier må være et multippel av $1/n!$, men slik at ikke ethvert multippel

nødvendigvis er oppnåelig avhengige av hvorvidt vi beregner like verdier av r_i (i = 1,2,3,...,n!), eller ikke.
La k bestå av mengden mulige multipler av $\frac{1}{n!}$. Hvis vi velger et signifikansnivå $\alpha = \frac{k}{n!}$, som er en av de oppnåelige p-verdiene vil ved en ensidig venstresidig test, når r* er den observerte Pearson's r:
P(type 1 feil) = P(p-verdi $\leq \alpha | H_0$) = P($\sum_{i=1}^{n!} I(r_i \leq r*) \leq k | H_0$) = $\frac{k}{n!} = \alpha$.
I tilfellene hvor vi har en oppnåelig p-verdi sier vi at testen er eksakt. Enkelte ganger har vi ikke tilgang på ønsket signifikansnivå α, men velger vi en lavere og oppnåelig verdi på α, er vi på den sikre siden, og vi sier at testen er konservativ.
Alternativt, i de tilfeller hvor den eneste hensikten er å forkaste, eller beholde nullhypotesen, kan vi sortere de n! beregnede Pearson's r verdiene og sjekke om den observerte Pearson's r-verdien ligger blant de 95% første (ensidig høyresidig test), 95% siste(ensidig venstresidig test), eller blant de 95% midterste verdiene i en tosidig test - hvis ikke dette er tilfelle forkaster vi nullhypotesen på et 5% signifikansnivå.
Nedsiden ved permutasjonstester er at det kan kreve mange og omfattende beregninger og at permutasjonstester primært brukes til å beregne p-verdier.

3.1.2 Tilpasning av den eksakte permutasjonsfordelingen til fordelingens momenter

Fordi hver målefordeling til Pearson r varierer i forhold til verdien på den faktiske populasjonskorrelasjonen ρ og målestørrelsen n er det uendelig målefordelinger av Pearson's r. Hvis alle disse målefordelingene til Pearson's r deler samme egenskaper vil det være enkelt for oss å gjøre statistisk inferens og på den måten si noe om hvor sannsynlig det vil være å observere en verdi av Pearson's r fra en måling gitt målestørrelse n og en populasjonskorrelasjonskoeffisient ρ. Vi vil se nedenfor, at under nullhypotesen $\rho = 0$ og under visse antagelser relatert til tredje og fjerde moment, at permutasjonsfordelingen kan tilnærmes med t-fordelingen.
Med $r = \frac{\frac{1}{n}\sum_i X_i Y_i - \bar{X}\bar{Y}}{s_X s_Y}$ (der s_x og s_y er delt på n) kan vi selvsagt oppdage den eksakte fordelingen til r ved å finne alle n! mulige ordninger av x'ene og y'ene og så studere fordelingen til r på bakgrunn av disse n! datapunktene jfr. 3.1.1 på forrige side. Når n er stor er dette svært krevende (n=10 gir 3.628.800 kombinasjoner, n=20 gir 2,4 milliarder milliarder kombinasjoner). For høye n approksimerer vi derfor heller den eksakte fordelingen ved å tilpasse fordelingen til dens momenter. Vi holder **y'ene faste** og permuterer x'ene og finner at jfr. [3, s. 492-493] :

$$E(\sum_i x_i y_i) = \sum_i y_i E(x_i) = \sum y_i \bar{x} = n\bar{x}\bar{y} \tag{63}$$

Likningen over impliserer at E(r)=0.
Pearson's r er invariant under lokaliseringsendringer i x og y så for enkelthetsskyld måler vi nå fra $\bar{x}$ og $\bar{y}$. Vi har da at:

$$var(\sum_i x_i y_i) = \sum_i y_i^2 var x_i + \sum_{i \neq j} \sum y_i y_j cov(x_i, x_j) \tag{64}$$

$$= \sum_i y_i^2 s_x^2 + \sum_{i \neq j} \sum y_i y_j \frac{1}{n(n-1)} \sum_{i \neq j} \sum x_i x_j \tag{65}$$

$$= n s_y^2 s_x^2 + ((\sum_i y_i)^2 - \sum_i y_i^2) \frac{1}{n(n-1)} ((\sum_i x_i)^2 - \sum_i x_i^2) \tag{66}$$

$$= n s_y^2 s_x^2 + \frac{n s_y^2 s_x^2}{n-1} \tag{67}$$

$$= \frac{n^2 s_y^2 s_x^2}{n-1} \tag{68}$$

Likningen over impliserer at $Var(r) = E(r^2) = (n^2 s_y^2 s_x^2)^{-1} var(\sum_i x_i y_i) = \frac{1}{n-1}$
Vi konstaterer at de to første momentene er uavhengig av de faktisk observerte verdier.
Vi kan tilsvarende finne at:

$$E(r^3) = \frac{n-2}{n(n-1)^2} \frac{k_3}{k_2^{3/2}} \frac{k_3'}{(k_2')^{3/2}} \tag{69}$$

$$E(r^4) = \frac{3}{n^2-1} (1 + \frac{(n-2)(n-3)}{3n(n-1)^2} \frac{k_4}{k_2^2} \frac{k_4'}{(k_2')^2}) \tag{70}$$

, der k'ene og k''ene er k-observatorene(jfr. [2, s. 281]) for henholdsvis x'ene og y'ene.
Hvis vi ser bort fra forskjellene på k-observatorene og målekumulantene([2, s. 281] har vi:

$$E(r^3) \approx \frac{n-2}{n(n-1)^2} g_1 g_1' \tag{71}$$

$$E(r^4) \approx \frac{3}{n^2-1} (1 + \frac{(n-2)(n-3)}{3n(n-1)^2} g_2 g_2') \tag{72}$$

, der g_1 og g_2 er skjevhet og kurtosis for x'ene(beregnet på utvalget) og g_1' og g_2' er skjevhet og kurtosis for y'ene. Vi har følgende uttrykk for utvalgskurtosisen (g_2) og utvalgsskjevheten(g_1):

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2)^2} - 3$$

$$g_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2)^{\frac{3}{2}}}$$

, som erstatter henholdsvis:

$$G_2 = \frac{k_4}{k_2^2} \tag{73}$$

$$= \frac{n^2((n+1)m_4 - 3(n-1)m_2^2)}{(n-1)(n-2)(n-3)} \frac{(n-1)^2}{n^2 m_2^2} \tag{74}$$

$$= \frac{n-1}{(n-2)(n-3)}((n+1)\frac{m_4}{m_2^2} - 3(n-1)) \tag{75}$$

$$= \frac{n-1}{(n-2)(n-3)}((n+1)g_2 + 6) \tag{76}$$

$$= \frac{(n+1)n(n-1)}{(n-2)(n-3)} \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} - 3\frac{(n-1)^2}{(n-2)(n-3)} \tag{77}$$

$$= \frac{(n+1)n}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{k_2^2} - 3\frac{(n-1)^2}{(n-2)(n-3)} \tag{78}$$

og

$$G_1 = \frac{k_3}{k_2^{3/2}} = \frac{n^2}{(n-1)(n-2)} \frac{m_3}{s^3}$$

,hvor $m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r$, for r=2, 3 og 4, er henholdsvis andre, tredje og fjerde utvalgsmoment rundt gjenommsnittet $\bar{x}$ jfr. [2, s. 229] og der $k_2 = \frac{n}{n-1}m_2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, $k_3 = \frac{n^2}{(n-1)(n-2)}m_3$ og $k_4 = \frac{n^2}{(n-1)(n-2)(n-3)}((n+1)m_4 - 3(n-1)m_2^2)$, som er de symmetriske og forventningsrette estimatorene til andre, tredje og fjerde kumulant uttrykt ved hjelp av utvalgsmomentene jfr. [2, s. 281].
Hvis $g_1, g_2, g_1^{'}$ og $g_2^{'}$ er konstante har vi at:

$$E(r^3) = O(n^{-2}) \tag{79}$$

$$E(r^4) = \frac{3}{n^2 - 1}(1 + O(n^{-1})) \tag{80}$$

og dermed vil når n går mot inf:

$$E(r^3) = 0 \tag{81}$$

$$E(r^4) = \frac{3}{n^2 - 1} \tag{82}$$

Vi konstaterer at momentene er presist de samme som momentene til den eksakte fordelingen til r når x,y kommer fra en bivariat normalfordeling med $\rho = 0$, som f.eks. kan vises ved hjelp av likning 28 på side 24. En god approksimasjon for permutasjonsfordelingen til r når vi antar at nullhypotesen $\rho = 0$ holder er altså:

$$dF = \frac{1}{B(\frac{1}{2}, \frac{1}{2}(n-2))}(1-r^2)^{\frac{1}{2}(n-4)}dr, -1 \leq r \leq 1 \tag{83}$$

og vi kan dermed bruke denne fordelingen eller ekvivalent at

$$t = (\frac{(n-2)r^2}{1-r^2})^{\frac{1}{2}} \tag{84}$$

har en Student's fordeling med (n-2) frihetsgrader til å utføre tester på r. Fordelingen er faktisk svært nøyaktig også for små n (når x,y ikke kommer fra en bivariat normalfordeling ved testing av ρ=0), som følger av at de to første momentene, både for den eksakte fordelingen for r der x,y kommer fra en bivariat normalfordeling, og for permutasjonsfordelingen er identiske. Adekvansen av approksimasjonen av tredje og fjerde moment, for permutasjonsfordelingen til r, avhenger selvsagt av verdiene til $g_1, g_1^{'}, g_2 og g_2^{'}$ men disse vil tendere mot å være små hvis F(x,y) er nær normal. Jo mindre g_1, g_1', g_2 og g_2' er jo mer robust blir approksimasjonen til permutasjonsfordelingen ovenfor når vi undersøker lineær sammenheng mellom de to variablene X og Y. Vi kan med formlene ovenfor lett regne ut g_1, g_2, g_1' og g_2'.
Konvergensen av permutasjonsfordelingen og normalteorifordelingen til en felles normalfordeling er blitt bevist av Hoeffding jfr. [14].

3.1.3 Tilnærmet permutasjonstest

Vi kan utføre en tilnærmet permutasjonstest(ofte kalt Monte Carlo permutasjonstest, eller tilfeldig permutasjonstest) uten å se på alle mulige n! parringer og uten å se på fordelingsegenskapene til r. Spesielt kan dette være interessant hvis F(X,Y) ikke er nær normal slik at g'ene typisk blir høye og n er for høy til å studere alle mulige parringer.
Sammenlignet med den ordinære permutasjonstesten plukker vi ut en viss andel, som blir tilfeldig valgt, av alle de mulige n! mulige utfall for den opprinnelige permutasjonstesten. Vi gjør som følger:
1) Med utgangspunkt i de originale parrede dataene (x_i, y_i) lager vi nye datasett (x_i, $y_{i'}$) der i' er en permutasjon av settet (1,2,..,n-1,n). Permutasjonen

i' velges tilfeldig, med lik sannsynlighet plassert på alle n! mulige permutasjoner. Vi trekker altså i' tilfeldig uten tilbakelegging fra settet 1,2,....,n-1,n.
2)Konstruerer korrelasjonskoeffisienten r fra de randomiserte data.
3)Utfør så 1) og 2) et stort antall ganger.
P-verdien for permutasjonstesten er da andelen av r-verdiene generert i step 2, som er $\geq/\leq$(ensidig høyresidig test/ensidig venstresidig test) enn den observerte Pearson's r, som ble kalkulert fra de originale data. Vi forkaster nullhypotesen om at X og Y er uavhengige dersom p-verdien overstiger et gitt α-nivå typisk på 0.05.
Det må bemerkes at p-verdien nå ikke er 100% korrekt, fordi vi ikke ser på alle de n! permutasjonene. Etter Np tilfeldige permutasjoner er det dog mulig å oppdage et konfidensintervall for p-verdien basert på den binomiske fordelingen. F.eks. hvis vi med Np=10000 tilfeldige permutasjoner estimerer p-verdien ved en ensidig venstresidig test til $\hat{p}$=0.05 vil et 99% konfidensintervall for den sanne p-verdien være (0.044,0.056) og dette må vi altså ta høyde for ved rapportering av p-verdier ved bruk av tilnærmede permutasjonstester. Med p=0.044 blir nedre grense nr. 440 i den sorterte fordelingen til de permuterte $\hat{\rho}$'ene og med p=0.056 blir nedre grense på nr. 560 i den sorterte fordelingen til de permuterte $\hat{\rho}$'ene.Vi kan selvsagt få konfidensintervallet for p så lite vi ønsker ved å øke Np.

3.2 Bootstrapping

Resultatene og teorien referert i seksjonen her er i all hovedsak hentet direkte fra [45].
Bootstrapping er et robust alternativ til inferens basert på parametriske antagelser når disse parametriske antagelsene er usikre, eller når parametrisk inferens er umulig, eller krever kompliserte formler for kalkulering av standardfeil.
Anta at vi har observert $(x_1, x_2, ...x_n) \sim$ F. Vi ønsker så å estimere en ukjent parameter $\hat{\theta} = \hat{\theta}(\mathbf{x})$, der $(\mathbf{x}) = (x_1, x_2, ...x_n)$. Spørsmål av interesse er:
1) er $\hat{\theta}$ forventningsrett, og hvis ikke hva er skjevheten($\beta_{\hat{\theta}}$) til estimatet?
2) hva er usikkerheten til $\hat{\theta}(\sigma_{\hat{\theta}})$
3) hvordan kan vi lage et konfidensintervall for θ.
Spørsmålene er knyttet til fordelingsegenskapene til $\hat{\theta}$, som avhenger av F. Et problem her er at F er ukjent noe som medfører at også fordelingsegenskapene til $\hat{\theta}$ er ukjente. Vi må gjøre tilnærminger av F for å svare på spørsmålene. Bootstrapping er da et alternativ. Ideeen bak bootstrapping er enkel (se f.eks. [45] jfr. [42]. Vi bruker et estimat $\hat{F}$ for F (siden vi ikke kjenner F) og ser så på egenskapene til $\hat{\theta}$ under $\hat{F}$. Det er to hovedvalg(A og B) med hensyn til hvordan vi estimerer F:

A) En mulighet er å anta at F tilhører en klasse av fordelinger beskrevet ved en eller flere parametere. Vi beskriver ofte da F med F_η, der η er en vektor av parametere, som beskriver fordelingen. Et estimat for F oppnås da ved å bruke estimater på de ukjente parameterne slik at $\hat{F} = F_{\hat{\eta}}$. Et eksempel kan være å anta normalfordelingen med forventning μ og standardavvik σ. F kan da estimeres ved å innsette maksimum likelihood estimatene $\hat{\mu}$ og $\hat{\sigma}$ for μ og σ. Bootstrapping basert på slike antagelser kalles for parametrisk bootstrapping. Devore and Berk behandler dette på side 339-340.
B)Et alternativ til parametrisk bootstrapping omhandles delvis i Devore and Berk [43, avsnitt 8.5]. Ideen er å gjøre minimale antagelser på F i utgangspunktet. Et mulig estimat for F i det tilfellet er den empirisk kumulative fordelingsfunksjonen som er gitt ved

$$\hat{F}(x) = \frac{1}{n}(\#x_i \leq x)$$

Merk at under $\hat{F}$ er $Pr(X = x_i) = \frac{1}{n}$ for i=1,2,...,n. og at trekking fra $\hat{F}$ svarer til å trekke fra $(x_1, x_2, ..., x_n)$ med tilbakelegging(merk forskjellen fra permutasjonstester). Dette er en viktig egenskap ved $\hat{F}$ som vi vil utnytte når vi skal utføre de nødvendige beregninger involvert i bootstrapping. Parameteren θ kan estimeres ved $\hat{\theta} = \hat{\theta}(\mathbf{x})$, en funksjon av data. Egenskapene til denne estimatoren er av interesse og spesielt er det av interesse å se på forventningsskjevheten til estimatoren definert ved $\beta_{\hat{\theta}} = E^F(\hat{\theta}(\mathbf{X})) - \theta(F)$. Problemet med å beregne forventningsskjevheten er at F er ukjent. Estimatet for forventningsskjevheten finner vi ved å erstatte den ukjente F med et estimat, som gir oss

$$E^{\hat{F}}(\hat{\theta}(\mathbf{X})) - \theta(\hat{F})$$

der vi med $\theta(\hat{F})$ mener den tilsvarende egenskap i $\hat{F}$ fordelingen, mens $E^{\hat{F}}(\hat{\theta}(\mathbf{X}))$ er forventningen til $\hat{\theta}(\mathbf{X})$ når $X_1, ..., X_n$ er uttrekk fra fordelingen $\hat{F}$.
Det største problemet med estimatet $E^{\hat{F}}(\hat{\theta}(\mathbf{X})) - \theta(\hat{F})$ er beregningen av $E^{\hat{F}}(\hat{\theta}(\mathbf{X}))$. Det er mulig å beregne denne siden $\hat{F}$ er kjent, men i praksis blir det fort et stort regnestykke. Merk at $\hat{F}$ er en diskret fordeling med n mulige utfall. For hele vektoren $\mathbf{x} = (x_1, x_2, ..., x_n)$ blir det dermed n^n mulige utfall, et stort tall selv for ganske lave n(n=5, n=6 og n=7 gir henholdsvis 3125, 46656, 823543 mulige utfall). Et alternativ da er å benytte oss av at uttrykket vi ønsker å beregne er en forventning og forventninger kan vi estimere ved hjelp av et utvalg fra den aktuelle fordelingen ved et såkalt bootstrapestimat. Lar vi $\mathbf{x}_1^*, ..., \mathbf{x}_B^*$ være B uavhengige uttrekk hver av størrelse n fra fordelingen

$\hat{F}$ og la $\theta_b^* = \hat{\theta}(\mathbf{x}_b^*)$, b=1,2,..,B. Da kan $E^{\hat{F}}(\hat{\theta}(\mathbf{X}))$ tilnærmes ved

$$E^{\hat{F}}(\hat{\theta}(\mathbf{X})) \approx E^{\hat{F}}(\hat{\theta}(\mathbf{X^*})) = \frac{1}{B}\sum_{b=1}^{B}\theta_b^*.$$

En slik tilnærming kalles ofte for Monte Carlo integrasjon og er et svært nyttig numerisk verktøy smln. forøvrig 3.1.3 på side 45. Bemerker at vi selv kan velge B og dermed hvor nøyaktig tilnærmingen vår skal være. Det endelige estimatet på forventningsskjevheten $\beta_{\hat{\theta}}$ blir:

$$b_{\hat{\theta}} = \frac{1}{B}\sum_{b=1}^{B}\theta_b^* - \theta(\hat{F}).$$

Vi konstaterer at det er to tilnærminger vi gjør ved beregning av dette estimatet. Den ene tilnærmingen skjer ved estimering av $E^{\hat{F}}(\hat{\theta}(\mathbf{X^*}))$. Den feilen vi her får vil ofte være neglisjerbar i forhold til den andre tilnærmingen vi gjør ved å erstatte F med $\hat{F}$.
Ovenfor har vi beskrevet hvordan forventningsskjevheten til en estimator $\hat{\theta}$ kan estimeres. Like enkelt kan andre egenskaper ved estimatoren analyseres. Anta f.eks. at vi er interessert i standardfeilen til estimatoren. Vi har:

$$\sigma_{\hat{\theta}} = \sqrt{E^F((\hat{\theta}(\mathbf{X}) - E^F(\hat{\theta}(\mathbf{X})))^2)}$$

med tilhørende bootstrapestimat

$$\sigma_{\hat{\theta}} = \sqrt{E^{\hat{F}}((\hat{\theta}(\mathbf{X}^*) - E^{\hat{F}}(\hat{\theta}(\mathbf{X}^*)))^2)}$$

, også i dette tilfelle kan estimatet i prinsippet beregnes analytisk, men i praksis vil vi tilnærme estimatet ved Monte Carlo integrasjon:

$$s_{\hat{\theta}} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\theta_b^* - \bar{\theta}^*)^2}$$

3.2.1 Bootstrapping og korrelasjonskoeffisienten ρ

Vi har at det naturlige estimatet for korrelasjonen ρ mellom to målinger av to variable X og Y er den empiriske korrelasjonen definert ved:

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2}}$$

Når vi skal lage bootstrap simuleringer av $\hat{\rho}$ må vi først lage et estimat $\hat{F}$ for denne bivariate sannsynlighetsfordelingen F. Vi skiller igjen på ikke-parametrisk og parametrisk bootstrapping.

Ikke-parametrisk bootstrapping Vi lar $Pr(X = x_i, Y = y_i) = \frac{1}{n}$ for i =1,2,...,n.
Simulering av $\hat{p}$ og beregning av ikke-parametriske bootstrap konfidensintervaller kan i henhold til forutsetningene ovenfor utføres i R ved at man:
1)trekker en tilfeldig indeks(med tilbakelegging) fra et utfallsrom, som har n^n mulige utfall før vi trekker tilhørende n par fra X,Y.
2)på bakgrunn av parene vi trakk i 1) lager vi et bootstrap- estimat $\hat{\rho}^*$ for $\hat{\rho}$.
3)vi gjentar så punktene i 1) og 2) f.eks B=1000 ganger slik at vi står igjen med 1000 bootstrap-estimater av $\hat{\rho}^*$. Vi kan så f.eks. lage et 95% ikke-parametrisk standard bootstrap konfidensintervall for ρ på følgende måte jfr [45, s. 8]:
4)sorter de beregnede bootstrapestimatene $\hat{\rho}^*$ i stigende rekkefølge.
5)Nedre grense for konfidensintervallet finner vi ved å plukke ut verdien lav, som er nummer 0.05/2*B blant de sorterte bootstrapestimaetene i 4) og så sette øvre grense til: $\hat{\rho}$ - (lav-$\hat{\rho}$).
Tilsvarende finner vi nedre grense ved å plukke ut verdien høy, som er nummer (1-0.05/2)*B blant de sorterte bootstrapestimatene i 4) og så sette nedre grense til: $\hat{\rho}$ - (høy-$\hat{\rho}$).
6)Dersom konfidensintervallet ikke inneholder 0 er dette en indikasjon på at det er en positiv sammenheng mellom variablene X og Y og jo lenger konfidensintervallet befinner seg vekk fra 0 jo sterkere er denne indikasjonen. Det bemerkes at bootstrapmetoden automatisk tar hensyn til den asymmetrien som er til stede slik at det ikke spiller noen rolle om bootstrapfordelingen er skjev eller ikke.
Hvis bootstrap-fordelingen til en estimator er symmetrisk($b_{\hat{\theta}} = 0$) kan man nøye seg med percentilkonfidensintervaller(lav,høy) i motsetning til de standard bootstrap konfidensintervallene, som tar hensyn til at **bootstrapfordelingen** kan være skjev.
Hvis vi, i tillegg til at bootstrapfordelingen er symmetrisk, antar at $\hat{p}$ er normalfordelt kan man finne tilsvarende konfidensintervaller, som i 5) ved: $(\hat{\rho} - 1,96 * s_{\hat{\rho}}, \hat{\rho} + 1,96 * s_{\hat{\rho}})$, der $s_{\hat{\rho}} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\rho_b^* - \bar{\rho}^*)^2}$. Dersom det er en negativ skjevhet($b_{\hat{\rho}}$ negativ) vil dette konfidensintervallet være litt forskjøvet nedover sammenlignet med bootstrapintervallet og motsatt hvis det er en positiv skjevhet ($b_{\hat{\rho}}$ positiv).

Parametrisk bootstrapping Jeg vil nå se på parametrisk bootstrapping relatert til ρ. I slike tilfeller antar vi at observasjonene stammer fra en kjent bivariat fordeling F_η. Dersom vi mener at X,Y kommer fra en bivariat normalfordeling består η av de fem parameterne innbyrdes korrelasjon, forventning og varians til hver av de to parameterne. Vi kan lage et 95% parametrisk

bootstrap konfidensintervall på følgende måte:
1) Estimer de fem parameterne E(X)=μ_X, E(Y)=μ_Y, Var(X)=σ_X, Var(Y)=σ_Y, Cor(X,Y)=$\rho_{X,Y}$ med $\hat{\eta}$ på vanlig måte fra utvalget vårt på n par av (X,Y).
2) Simuleringer fra $F_{\hat{\eta}}$ kan nå utføres ved å simulere n variable fra den tilpassede bivariate normalfordelingen. For å simulere slike variable kan vi benytte oss av følgende egenskaper ved den bivariate normalfordelingen [43, p. 255-256]: $X \sim N(\mu_X, \sigma_X)$ og $Y|X \sim N(\mu_Y + \frac{\rho(X-\mu_X)\sigma_Y}{\sigma_X}, \sigma_Y\sqrt{1-\rho^2})$. Vi kan da først simulere X fra dens marginale normalfordeling og deretter $Y|X$ fra den betingede normalfordeling innsatt parameterestimatene fra 1) og så beregne korrelasjonen ρ^* til de n simulerte parene (X^*, Y^*). Vi gjentar så dette f.eks B=1000 ganger.
3) Vi kan da f.eks. beregne bootstrapestimatene:
skjevhet = $b_{\hat{\rho}} = \bar{\rho^*} - \hat{\rho}$
standardavvik = $s_{\hat{\rho}} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\rho_b^* - \bar{\rho}^*)^2}$
4)Vi kan også beregne et 95 % standard bootstrap konfidensintervall. Nedre grense for konfidensintervallet finner vi ved å plukke ut verdien lav, som er nummer 0.05/2*B blant de sorterte bootstrapestimaetene for ρ^* i 2) og så sette øvre grense til: $\hat{\rho}$ - (lav-$\hat{\rho}$).
Tilsvarende finner vi nedre grense ved å plukke ut verdien høy, som er nummer (1-0.05/2)*B blant de sorterte bootstrapestimatene i 2) og så sette nedre grense til: $\hat{\rho}$ - (høy-$\hat{\rho}$).
Jeg bemerker at dersom vi kan anta en bivariat normalfordeling vil metoden her antagelig fungere bra og bør gi nogenlunde tilsvarende resulater som Fisher-transformasjonen og som den eksakte fordelingen til Pearson's r.
Hvis vi antar at $\hat{p}$ er normalfordelt kan man finne tilsvarende konfidensintervaller, som i 4) ved: $(\hat{\rho}-1,96*s_{\hat{\rho}}, \hat{\rho}+1,96*s_{\hat{\rho}})$, der $s_{\hat{\rho}} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\rho_b^* - \bar{\rho}^*)^2}$. Dersom det er en negativ skjevhet($b_{\hat{\rho}}$ negativ) vil dette konfidensintervallet være litt forskjøvet nedover sammenlignet med bootstrapintervallet og motsatt hvis det er en positiv skjevhet ($b_{\hat{\rho}}$ positiv).

Varsomhet ved bruk av bootstrapping Den grunnleggende motivasjonen for at bootstrapmetoden gir gode tilnærminger til de sanne egenskapene til estimatet $\hat{\theta}$ ligger i at $\hat{F}$ er en rimelig tilnærming til F. Dersom det er grunnlag for å bruke bootstrapmetoden, kan man vise at bootstrapestimatene, under gitte antagelser, er asymptotisk korrekte.
Vi har sett på standard bootstrapintervaller og kommentert bruk av percentilkonfidensintervaller og normaltilnærming, som mulige metoder ved kon-

struksjon av konfidensintervaller. Tilsvarende som for bootstrapmetodene er normaltilnærmingen også korrekt i asymptotisk forstand. Motivasjonen for å bruke mer kompliserte bootstrapmetodikker, som f.eks. standard bootstrap intervaller, er at disse stort sett vil være bedre på den måten at, for et endelig antall observasjoner, vil den feilen vi gjør under tilnærming, typisk være mindre enn den vi får ved en normaltilnærming.
Det eksisterer andre lignende og ulike typer bootstrap konfidensintervaller enn ovenfornevnte. Vi har 'studentized bootstrap intervals', 'bias bootstrap intervals', 'accelerated(standard) bootstrap intervals' og 'bias and accelerated bootstrap intervals' - metodene kan i gitte situasjoner være fruktbare og gi gode estimater. Bruk og utledning av disse metoder gjøres ikke her og det henvises f.eks. til [53] for innsikt i disse metoder.

4 Ikke-parametriske korrelasjonsmål

Mange av resultatene i seksjonen her er hentet fra [1].
Vi har i seksjon 2 på side 14 sett at det, når den underliggende fordeling til (X,Y) ikke er bivariat normalfordelt, ikke er korrekt å bruke den eksakte fordelingen til Pearson's r til å gjøre inferens relatert til ρ. Transformasjoner på data, som Fishertransformasjonen, blir dermed også uaktuelt.
Vi har også sett at når vi tilnærmer permutasjonsfordelingen til Pearson's r, via momentene, jfr. seksjon 3.1.2 på side 42 kan vi teste $H_0 : \rho = 0$ ved hjelp av t-fordelingen. Det er dog noe matematisk krevende å tilnærme permutasjonsfordelingen med t-fordelingen i tillegg til at tilnærmingen ikke alltid er god - spesielt gjelder dette når vi har lite data og jobber med fordelinger der skjevhet og kurtosis for X og/eller Y avviker klart fra normalfordelingen.
Vi konstaterer, på bakgrunn av ovenfornevnte, at i situasjoner med lite data må vi typisk bevege oss over i fordelingsfrie prosedyrer, som ikke krever mye data.
Vi har i seksjon 3 på side 40 sett at den eksakte fordelingen til Pearson's r, under $H_0 : \rho = 0$, via permutasjon kan være omstendelig å finne frem til. Årsaken er at den eksakte fordelingen til r avhenger av de faktiske observerte verdiene av variablene X og Y, som jo er tilfeldige variable,(og dermed gir opphav til uendelig med eksakte fordelinger for r) og når n er høy vil det i tillegg kreve svært mange permuteringer. Med dagens datakraft er det likevel enkelt og implementere disse beregningene i f.eks statistikkprogrammet R for å finne frem til den eksakte permutasjonsfordelingen når målestørrelsen er lav og spesielt når den er ≤ 10 slik den typisk er i oppgaven her.

Vi har også sett, jfr. seksjon 3 på side 40, at den fordelingsfrie prosedyren bootstrapping kan gi dårlige resultater når vi ikke mener å kunne legge fornuftige antagelser på F. I tilfelle med n=5 og n=6, slik som i oppgaven her, og ellers lite kunnskap om populasjonen, er det vanskelig å legge gode antagelser på F. I tillegg til tilnærmingen av F(og tilnærminger via Monte Carlo integrasjon) vil bootstrapping kunne være sensitiv for uteliggere(dersom bootstrapestimatene ρ^* er beregnet med flere uteliggere enn vår observerte $\hat{\rho}$). Av nevnte grunner vil bootstrapping ofte være et mindre godt alternativ til å teste $H_0 : \rho = 0$.
Generelt har vi også sett at et vesentlig drawback ved beregning av Pearson's r, spesielt for små målinger, er at selve korrelasjonen man beregner er svært sensitiv for uteliggere og det vil i mange praktiske situasjoner være nyttig å finne korrelasjonsmål, som ikke påvirkes av uteliggere.
Spørsmålet er om vi kan finne statistiske prosedyrer, som forblir gyldige for en vid klasse av populasjonsfordelinger - f.eks for alle kontinuerlige fordelinger!?. Prosedyrene kalles fordelingsfrie, fordi gyldigheten av bruken av metoden ikke på noen måte avhenger av underliggende fordeling på X,Y forutsatt at de er kontinuerlige. Hvis vi kan finne slike fordelingsfrie prosedyrer vil de også være gyldige for normalfordelinger og våre robusthets resulateter vil bli presise og forsikrede. Vi ønsker så å si og komme frem til enklere mål for korrelasjonskoeffisienten Pearson's r slik at det blir lettere å studere fordelingen til det alternative korrelasjonsmålet herunder gjøre inferens.
Et mål er altså å fjerne permutasjonsfordelingens avhengighet til de tilfeldig varierende observasjonene. Dette kan vi gjøre ved å rangere X og Y. F.eks definerer vi rangen til y_i som dets posisjon blant den ordnede observatoren til y slik at rang(y_i) = i. Rangeringer er invariante under monotone transformasjoner på variablene noe som forsterker tilbøyeligheten til å se på tester basert på rangeringer. En hvilken som helst monoton transformasjon av variablene vil opprettholde uavhengighetshypotesens invariante egenskap og rangeringer er derfor naturlige størrelser å bruke.
Rangeringen av variablene medfører typisk at vi må erstatte de observerte verdiene av (X,Y) med tilhørende korrelasjonskoeffisient Pearson's r med nye verdier (X',Y') med korrelasjonskoeffisient r' slik at permutasjonsfordelingen til r' er den samme for hver måling selvom r' vil variere fra måling til måling. Det er m.a.o ingen tilfeldighet i de observerte rangerte data, som det er, i de observerte data som ikke er rangert.
Ved å benytte rangeringer kan korrelasjonsmålene også benyttes på ordinale data i tillegg til kontinuerlige data.
Jeg vil nå se på ulike korrelasjonskoeffisienter, som bygger på rangerte data.

4.1 Spearman's rang korrelasjonskoeffisient r_s - et estimat for populasjonskorrelasjonskoeffisienten ρ_s

Vi definerer:

$$r_s = 1 - \frac{6\sum d^2}{(n^3 - n)} \tag{85}$$

, der d er definert under seksjon 4.1.1. r_s kalles gjerne Spearman's rang korrelasjonskoeffisient etter den dyktige psykologen Charles Spearman, som først introduserte korrelasjonsmålet i 1906, og da som et substitutt til den ordinære korrelasjonskoeffisienten r. Dog skriver K.Pearson i hans biografi av Galton at Galton jobbet med rangkorrelasjoner før han så på korrelasjon mellom kontinuerlige variable rundt 1875, men Galton publiserte ingenting eksplisitt. Opphavet til r_s er derfor noe uklart.

4.1.1 Spearman's r_s et spesialtilfelle av den generaliserte korrelasjonskoeffisienten

Setter $a_{ij} = p_j - p_i$, $b_{ij} = q_j - q_i$ og $d = p_i - q_i$, der p_i er rangen til det i'te medlem relatert til de observerte x-verdiene/kvalitetene og q_i er rangen til det i'te medlemmet relatert til de observerte y-verdiene/kvalitetene. Verdier når den observerte variabelen er kontinuerlig og kvaliteter når den observerte variabelen er ordinal. Både p_i og q_i går fra 1 til n og slik at $\sum_{i,j}(p_j - p_i)^2$ er lik $\sum_{i,j}(q_j - q_i)^2$ og vi får at den generelle korrelasjonskoeffisienten jfr. likning 9 på side 16 tar formen :

$$\Gamma = \frac{\sum_{i,j}(p_j - p_i)(q_j - q_i)}{\sum_{i,j}(p_j - p_i)^2} \tag{86}$$

Vi har at:

$$\sum_{i,j}(p_j - p_i)(q_j - q_i) = \sum_{i=1}^{n}\sum_{j=1}^{n}(p_i q_i) + \sum_{i=1}^{n}\sum_{j=1}^{n}(p_j q_j) - \sum_{i=1}^{n}\sum_{j=1}^{n}(p_i q_j + p_j q_i) = \tag{87}$$

$$2n\sum_{i=1}^{n} p_i q_i - 2\sum_{i=1}^{n} p_i \sum_{j=1}^{n} q_j = 2n\sum_{i=1}^{n} p_i q_i - \frac{1}{2}n^2(n+1)^2, \tag{88}$$

,siden både $\sum p_i$ og $\sum q_j$ er summen av de første n positive tallene og der den aritmetiske rekken 1 + 2 + ...+ n summeres til $\frac{1}{2}n(n+1)$

Jeg setter videre:

$$\sum d^2 = \sum_i^n (p_i - q_i)^2 = 2\sum_i^n p_i^2 - 2\sum_i^n p_i q_i \tag{89}$$

, som gir at uttrykket ovenfor kan skrives som:

$$\sum_{i,j}(p_j - p_i)(q_j - q_i) = 2n\sum_i^n p_i^2 - \frac{1}{2}n^2(n+1)^2 - n\sum d^2, \tag{90}$$

,der $\sum_i^n p_i^2$ er summen av kvadratene av de første n positive hele tallene lik $\frac{1}{6}n(n+1)(2n+1)$ og dermed blir $\sum_{i,j}(p_j - p_i)(q_j - q_i) = \frac{1}{6}n^2(n^2-1) - n\sum d^2$
Videre har vi at:

$$\sum_{i,j}(p_j - p_i)^2 = 2n\sum_i^n p_i^2 - 2\sum_{i,j}^n p_i p_j = 2n\sum_i^n p_i^2 - 2(\sum_i^n p_i)^2 \tag{91}$$

$$= 2n\frac{1}{6}n(n+1)(2n+1) - 2(\frac{1}{2}n(n+1))^2 = \frac{1}{6}n^2(n^2-1) \tag{92}$$

Vi får dermed at:

$$\Gamma = \frac{\sum_{i,j}(p_j - p_i)(q_j - q_i)}{\sum_{i,j}(p_j - p_i)^2} = 1 - \frac{6\sum d^2}{n^3 - n} \tag{93}$$

, som er definisjonen på Spearman's r_s

4.2 Kendall's rang korrelasjonskoeffisient t - et estimat for populasjonskorrelasjonskoeffisienten τ

Kendall's t har blitt diskutert av mange statistikere tidligere, som f.eks Fechner og Lipps rundt 1900, av Lindeberg i 1920 og historiske detaljer er gitt av Kruskal i 1958, men korrelasjonskoeffisienten Kendall's t ble først allment brukt etter at en serie av artikler av M.G.Kendall ble utgitt f.o.m 1938 og konsolidert i en monografi av Kendall i 1962, som også er blitt en referanse hva gjelder spørsmål om bruk av Kendall's t og r_s, som mål på korrelasjon.

4.2.1 Kendall's t sett på som en koeffisient av overensstemmelse(concordance)

Blant de n rangerte parene ser vi på to par f.eks (x_i, y_i) og (x_j,y_j) og definerer dem som overensstemmende(concordant) hvis $y_i < y_j$ når $x_i < x_j$, eller $y_i > y_j$ når $x_i > x_j$, som begge er ekvivalente med at (x_i - x_j)(y_i - y_j) > 0.

Tabell 6: Rate of personal effort i klassisk og skøyting - kontinuerlig skala.

Utøver	A	B	C	D	E	F
RPE Klassisk(X)	11	8	14	5	16	13
RPE Skøyting(Y)	7	15	12	9	10	17

Tabell 7: Rate of personal effort i klassisk og skøyting - ordinal skala 1

Utøver	D	B	A	F	C	E
RPE Klassisk(X)	1	2	3	4	5	6
RPE Skøyting(Y)	2	5	1	6	4	3

Vi definerer de to parene som uoverensstemmende(discordant) hvis $y_i < y_j$ når $x_i > x_j$, eller $y_i > y_j$ når $x_i < x_j$, som begge er ekvivalent med at (x_i - x_j)(y_i - y_j) < 0. Ved beregningen av Kendall's t vil vært par med endelig score lik 1 kalles overensstemmende og hvert par med score -1 kalles uoverensstemmende. Vi lar P være antall overensstemmende par og Q antall uoverensstemmende par og S = P-Q er antall overensstemmende par som overstiger antall uoverensstemmende par. P og Q er altså totalene av henholdsvis de positive og negative scorene slik at P+Q = $\frac{1}{2}n(n-1)$. Vi får dermed at Kendall's t kan skrives på følgende former:

$$t = \frac{S}{\frac{1}{2}n(n-1)} = \frac{2S}{n(n-1)} = \frac{2(P-Q)}{n(n-1)} = 1 - \frac{4Q}{n(n-1)} = \frac{4P}{n(n-1)} - 1. \tag{94}$$

Siden totalt antall par er $\frac{1}{2}n(n-1)$ ser vi at Kendall's t er definert som andelen av overensstemmende par minus andelen av uoverensstemmende par og Kendall's t er derfor et relativt mål av overensstemmelse mellom de to settene, relatert til henholdsvis X og Y, av rangeringer. Vi ser på et eksempel: Antar at n=6 idrettsutøvere er rangert i forhold til deres egenfølelse(RPE=rate personal effort) av prestasjon i henholdsvis klassisk og skøyting, jfr. tabell 6. Vi ønsker å se om det er noen relasjon mellom skøyting og klassisk og rerangerer klassiskdataene i naturlig rekkefølge(der naturlig rekkefølge er 1,2,3...) for å få et bedre overblikk på data jfr. tabell 7. Vi konstaterer at sammenhengen mellom rangeringene for henholdsvis klassisk og skøyting er langt fra perfekt. Studerer så alle par av objekter. For paret AB i klassisk ser vi at rangeringene til A og B er henholdsvis 3 og 2. Rangeringene 3 og 2 er ikke i naturlig rekkefølge, jfr. tabell 6, og vi gir dette paret sco-

ren -1. For samme paret i skøyting er rangeringene 1 og 5, som er i naturlig rekkefølge, jfr. tabell 6 på forrige side, og vi gir derfor dette paret scoren 1. Produktet av scorene for paret AB er -1, som er endelig score for paret AB. Vi gjør så dette for samtlige par og ender med følgende scores:
AB = -1
AC = 1
AD = -1
AE = 1
AF = 1
BC = -1
BD = 1
BE = -1
BF = 1
CD = 1
CE = -1
CF = -1
DE = 1
DF = 1
EF = -1
Vi fikk totalt P = 8 i positiv og score Q = 7 i negativ score. Samlet score S = 8 - 7 = 1. Totalt mulig score er $\frac{1}{2}n(n-1) = \frac{1}{2} * 6 * 5 = 15$. Vi får dermed:

$$t = \frac{1}{\frac{1}{2}6(6-1)} = 1/15 \approx 0.0667. \tag{95}$$

, som tyder på at det er vanskelig å ha en god egenfølelse knyttet til prestasjon både i skøyting og i klassisk. Hvis det er full sammenheng mellom rangeringene er Q=0 og $t = 1$ og tilsvarende er $t = -1$ når P = 0. Hvis P = Q er $t = 0$.
Bemerker at det er mange shortcuts for å definere S eller ekvivalent P og Q på - for innsyn i disse metoder ber jeg leseren om og f.eks. slå opp i [1, s. 5].

4.2.2 Kendall's t sett på som en koeffisient av uorden(disarray)

Anta at vi har rangert X. Vi tar så og transformerer rangeringen av Y ved suksessivt å bytte nabopar. Minimalt antall byttinger for at Y skal bli rangert likt som X kaller vi s. Det viser seg at $s = Q = \frac{1}{2}(\frac{1}{2}n(n-1) - S)$, som gir en enkel sammenheng mellom minste antall byttinger s og antall negative scores Q eller total score S.
Illustrerer dette med utgangspunkt i eksempelet i seksjon 4.2.1 på side 54 hvor variabelen klassisk(X) er ordnet slik at verdiene kommer i naturlig

Tabell 8: Rate of personal effort i skøyting og klassisk - ordinal skala 2

Utøver	D	B	A	F	C	E
RPE Klassisk(X)	1	2	3	4	5	6
RPE Skøyting(Y)	2	5	1	6	4	3
d	-1	-3	2	-2	1	3

rekkefølge jfr. tabell 6 på side 55 og tabell 7 på side 55. Vi transformerer så med enkeltvise byttinger slik at skøytevariabelen (Y) også er ordnet slik at verdiene kommer i naturlig rekkefølge. For å få den laveste verdien 1 til skøytevariabelen først, må jeg først bytte om 5 og 1 for deretter og bytte 1 og 2. Vi har da følgende sekvens etter to byttinger:
RPE Skøyting(Y) 1 2 5 6 4 3
Konstaterer at 2 allerede står på rett plass, men 3 må bytte plass først med 4, så med 6 og så med 5 for å komme på rett plass - totalt 3 byttinger må til for å få 3 på rett plass. Vi står nå med sekvensen:
RPE Skøyting(Y) 1 2 3 5 6 4
Vi må nå bare bytte 4 først med 6 og så med 5 altså to byttinger for å få skøytevariabelen i naturlig rekkefølge og slik at vi ender med sekvensen:
RPE Skøyting(Y) 1 2 3 4 5 6
For å få skøytevariabelen rangert i naturlig rekkefølge har vi måttet gjøre totalt s = 2+3+2 = 7 = Q = 0.5*(0.5*6*5 - 1) = 7 byttinger. Dette er minste antall byttinger, som gjør at vi får transformert skøytevariabelen i naturlig rekkefølge. Man kan vise at dette minste antall byttinger er gitt ved: s = Q $\Leftrightarrow$ $s = \frac{1}{2}(\frac{1}{2}n(n-1) - S)$ jfr. [1, s. 30]
Scoren til Q er altså antall par som ikke er i naturlig rekkefølge. Et par som ikke er i rett rekkefølge kaller vi en inversjon og Kendall's t er altså bare en lineær funksjon av antall inversjoner($t = 1 - \frac{4s}{n(n-1)}$).
Man kan også oppfatte r_s, som en inversjonskoeffisient der hver inversjon blir vektet. Hvis f.eks. paret med rangeringer (i,j) inverteres slik at (i<j) og vi gir scoren (j-i) for enhver slik inversjon og vi deretter summerer alle disse scorene for å oppdage en totalscore V har vi at V = $\sum \frac{d^2}{2}$ og dermed at r_s = 1 - $\frac{12V}{n^3-n}$. For et generelt bevis se [1, s. 31].
Illustererer nå sammenhengen mellom V og d^2 med et eksempel der vi fortsatt ser på eksempelet i tabell 6 på side 55 og tabell 7 på side 55, men slik at vi også tar med differensen d jfr. tabell 8. Rangeringene som må inverteres er:
Ranger Vekt
2 1 1
5 1 4

5 4 1
5 3 2
6 4 2
6 3 3
4 3 1
Dette gir V = 1 + 4 + 1 + 2 + 2 + 3 + 1 = 14 og $\sum d^2/2 = ((-1)^2 + (-3)^2 + 2^2 + (-2)^2 + 1^2 + 3^2)/2 = 14$.
Vi konstaterer sammenhengen og ekvivalensen mellom totalt antall vektede inverteringer og $\sum d^2/2$.
Legg også merke til at vi har 7 inverteringer for å få skøytedataene i samme orden som klassisk dataene og altså er Q=7.
I dette eksempelet blir:
$t = 1 - \frac{4*7}{6*5} = 0.0667$
$r_s = 1 - \frac{12*14}{6^3-6} = 1 - 0.8 = 0.2$
og vi konstaterer en langt høyere korrelasjon med r_s enn ved bruken av Kendall's t.

4.2.3 Kendall's t - et spesialtilfelle av den generelle korrelasjonskoeffisienten

La p_i være rangen til det i'te objektet og p_j rangen til det j'te objektet der begge rangeringene viser til X-variabelen. Anta at vi gir scoren +1 når $p_j > p_i$ og scoren -1 hvis hvis $p_j < p_i$. Da blir $a_{ij} = +1$ når $p_i < p_j$ og $a_{ij} =$ -1 når $p_i > p_j$. Gjør vi tilsvarende for b og Y-scorene får vi at den generelle korrelasjonskoefisienten jfr. likning 9 på side 16 og likning 86 på side 53 blir:

$$\Gamma = \frac{\sum a_{ij}b_{ij}}{\sqrt{\sum a_{ij}^2 \sum b_{ij}^2}} = \frac{2S}{n(n-1)} \tag{96}$$

, siden $\sum a_{ij}b_{ij}$ er lik 2 ganger summen S(to fordi et hvilket som helst par forekommer to ganger- både som (i,j) og (j,i) i summeringen). Videre er $\sum a_{ij}^2$ bare antall termer a_{ij}, som jo er n(n-1) og helt tilsvarende blir det for $\sum b_{ij}^2$.

4.3 Like rangeringer (Tied ranks)

Hvis man ikke klarer å skille to eller flere individer/observasjoner, på en slik måte at man kan gi dem ulik rang, har individene/observasjonene samme rang(tied). Spørsmålet er da hvilken rang vi skal gi de likt rangerte individene/observasjonene. Det eksisterer ulike metoder å omgå problemet med likt rangerte observasjoner/individer på. Vi skal bruke den såkalte gjennomsnittrangerings-metoden, som illustreres med følgende eksempel.

Dersom f.eks observasjonene, som skal fordeles på posisjonene 3,4 og 5 ikke kan skilles på kan man gi hver av dem rangeringen $\frac{1}{3}(3+4+5) = 4$, eller dersom man ikke klarer å skille på posisjonene 9 og 10 får de tilhørende observasjonne/individene hver seg rang 9.5. En følge av å benytte denne metoden er at summen av rangeringene forblir den samme, som om rangeringene ikke var like. Vi vil nå se på konsekvenser av likt rangerte observasjoner ved beregning av Kendall's t og r_s

4.3.1 Beregning av Kendall's t ved like rangeringer

I 4.2.3 på forrige side så vi at det, for hvert par, ble gitt en score på henholdsvis 1 eller -1 avhengig av om paret var rangert i naturlig rekkefølge eller ei. Når vi ikke klarer å skille på rangeringene gir vi derimot scoren 0. Dette innebærer, som vi skjønner at nevner i likning 96 på forrige side ikke vil endre seg, men absoluttverdien av telleren vil ikke lenger kunne ta en så høy verdi, som nevneren, og dermed vil -1 < t < 1. For å omgå problemet med at vi ikke kan oppnå en korrelasjon på 1 eller -1 erstatter vi bare nevneren $\frac{1}{2}n(n-1)$ i seksjon 4.2.3 på forrige side med $\frac{1}{2}\sqrt{\sum a_{ij}^2 \sum b_{ij}^2}$, der a_{ij} er scoren til det i'te og j'te medlemmmet av variabelen X og b_{ij} er scoren til det i'te og j'te medlemmmet av variabelen Y. Hvis det ikke er likt rangerte observasjoner blir alle $a_{i,j}^2 = 1$ slik at $\sum a_{i,j}^2 = n(n-1)$ og tilsvarende for $b_{i,j}^2$ og $\sum b_{i,j}^2$. Dette impliserer videre at $\frac{1}{2}\sqrt{\sum a_{ij}^2 \sum b_{ij}^2} = \frac{1}{2}n(n-1)$, som forventet.
Hvis vi har u likt rangerte observasjoner så vil alle mulige u*(u-1) par ha score 0 og konsekvent får vi da at $\sum a_{i,j}^2 = n(n-1) - \sum u(u-1)$, der den siste summeringen går over alle samlinger av like rangeringer. Definerer så:

$$U = \frac{1}{2}\sum u(u-1) \tag{97}$$

, relatert til like rangeringer i variabel X og

$$V = \frac{1}{2}\sum v(v-1) \tag{98}$$

for like rangeringer i variabel Y og slik at vi får at en alternativ form av Kendall's t til å være:

$$t_b = \frac{S}{\sqrt{\frac{1}{2}n(n-1) - U}\sqrt{\frac{1}{2}n(n-1) - V}} \tag{99}$$

Vi ser at t_b også er et spesialtilfelle av den generelle korrelasjonskoeffisienten Γ jfr. seksjon 2.2 på side 16. Vi ser også, at dersom vi ikke gjør en

justering på nevneren ved beregning av Kendall's t-koeffisienten ved likt rangerte observasjoner, vil vi konsekvent få en lavere korrelasjon enn om vi gjør justeringen. Hvis vi ikke justerer nevneren, som forklart ovenfor, ved like rangeringer ,beregner vi den såkalte t_a-koeffisienten. t_a-koeffisienten sammenfaller selvsagt med Kendall's t når vi ikke har likt rangerte data.

Bruk av t_a eller t_b Fra utledningen av den generelle Γ-koeffisienten følger det at dersom det eksisterer likt rangerte observasjoner er det t_b, som er det korrekte korrelasjonsmålet mellom to variable. Likevel vil det enkelte ganger, der det eksisterer likt rangerte observasjoner, være mer forklarende å bruke t_a, som korrelasjonsmål. La meg Illustrere dette med et eksempel:
La oss anta at to observatører evaluerer prestasjon på 10 utøvere hvor begge observatørene rangerer dem etter egen oppfatning. En eller begge observatører rangerer noen av utøverne likt. Grunnen til at t_b er det beste korrelasjonsmålet er at det er enigheten mellom observatørene vi her måler - ikke nøyaktigheten av hvordan de treffer med en objektiv rangering, som f.eks fispunkter til løperne.
Anta så at det eksisterer en objektiv rekkefølge og at en observatør skal forsøke å rangere utøverne i henhold til denne objektive rekkefølgen. Formålet med rangeringen til observatøren er da å måle nøyaktigheten av rangeringen mot den objektive rangeringen. Formen på t_b vil da gi vekt til det faktum, dersom observatøren rangerer noen likt, at variasjonen i estimatene reduseres. Bruken av t_b tar da høyde for likt rangerte observasjoner selvom det faktisk ikke skal være likt rangerte observasjoner siden det eksisterer en faktisk objektiv rekkefølge. I disse tilfellene vil t_a ofte vise seg å være mer fornuftig da de som klarer å skille på rekkefølgen i større grad blir premiert med høyere korrelasjon av egen bedømmelse opp mot den objektive rangeringen.
Et annet interessant eksempel i denne sammenheng er som følger. Dersom man velger ut dommere til internasjonale konkurranser utfra hvordan de korrelerer med hverandre vil dette være svært uheldig siden det ikke nødvendigvis medfører at den objektivt sett beste prestasjonen blir premiert best. Man bør derfor velge ut dommere, dersom det er mulig, på bakgrunn av hvordan de hver seg korrelerer med en ojektiv rekkefølge, slik at de som best korrelerer med den objektive rekkkefølgen blir plukket ut. Ved beregningen av korrelasjonen for den enkelte dommer mot den objektive rekkefølge vil det jfr. foregående avsnitt kunne være fornuftig å velge korrelasjonsmålet t_a. Problemet med denne tilnærmingen er selvsagt at en objektiv rekkefølge ofte typisk er i mer eller mindre kontinuerlig forandring og at man må kunne definere den objektive rekkefølgen.

Sammenhengen mellom Kendall's t og t_a Anta at du ikke klarer å skille på et sett u likt rangerte data i variabel X. Dersom du beregner Kendall's t for alle mulige tenkelige rekkefølger av de likt rangerte observasjonene og tar gjennomsnittet av alle disse Kendall's t'ene vil vi få t_a beregnet for den opprinnelige rangeringen. Dette følger av at i de u! ulike kombinasjonene vil hvert par skje et likt antall ganger i AB-rekkefølge, som i BA-rekkefølge slik at allokering av score på 1 i et tilfelle blir -1 i det andre tilfelle, som er ekvivalent med å allokere 0 i gjennomsnitt.

4.3.2 Beregning av r_s ved like rangeringer

Ved beregning av Spearman's r_s møter vi på tilsvarende problemer ved likt rangerte observasjoner, som ved beregning av Kendall's t-koeffisienten. Vi opererer her med de tilsvarende uttrykkene r_{sa} og r_{sb} for henholdsvis t_a og t_b. Hvis vi fortsatt lar u være et sett av antall likt rangerte observasjoner for variabelen X og tilsvarende lar v være et sett av antall likt rangerte observasjoner for variabelen Y definerer vi:

$$U' = \frac{1}{12}\sum(u^3 - u) \tag{100}$$

$$V' = \frac{1}{12}\sum(v^3 - v) \tag{101}$$

, der summeringene går over alle sett av like rangeringer for henholdsvis variabel X og variabel Y.
Vi kan da vise jfr. [1, s. 43-46] at:

$$r_{sa} = 1 - \frac{6\sum d^2 + U' + V'}{n^3 - n} \tag{102}$$

og

$$r_{sb} = \frac{\frac{1}{6}(n^3 - n) - \sum d^2 - U' - V'}{\sqrt{(\frac{1}{6}(n^3 - n) - 2U')}\sqrt{(\frac{1}{6}(n^3 - n) - 2V')}} \tag{103}$$

Analogien til t_a og t_b Hvorvidt det er fornuftig å bruke r_{sa} eller r_{sb} er helt tilsvarende diskusjonen ovenfor om når det er fornuftig å bruke t_a eller t_b.
Videre er det slik at r_{sa} er gjennomsnittet av alle r_s-koeffisienter, som vi finner ved å erstatte de likt rangerte observasjonene med alle mulige tenkelige rekkefølger av disse likt rangerte observasjonene på samme måte som for t_a og Kendall's t, men med en noe annen forklaring jfr. [1, s. 46]

Tabell 9: VO2max for kvinner og menn - ordinal og nominal skala

Rangering VO2max	1	2	3	4	5	6	7	8	9
Kjønn	K	M	K	K	M	K	M	M	M

Tabell 10: VO2max for kvinner og menn - ordinal skala

Rangering VO2max	1	2	3	4	5	6	7	8	9
Rangering Kjønn	2.5	7	2.5	2.5	7	2.5	7	7	7

4.3.3 Like rangeringer herunder en binær variabel

Enkelte ganger er det ønskelig å måle sammenhengen mellom to egenskaper, der den ene egenskapen f.eks bygger på rangering og der den andre egenskapen er en klassifisering av hvorvidt individet/observasjonen faller i den ene eller andre kategori. F.eks ønsker vi å se om det er noen sammenheng mellom VO2max og kvinner og menn jfr. tabell 9. Dersom vi antar at inndeling i kjønn i seg selv er en rangering vil vi for de første 4 jentene ha en gjennomsnittelig rang på 2.5 og 7 for de neste fem mennene - se tabell 10. Vi får da at S = 5 + -3 + 4 + 4 + -1 + 3 + 0 + 0 + 0 = 12, U = 0(ingen likt rangerte verdier i X) og V = $\frac{1}{2} * 4 * 3 + \frac{1}{2} * 5 * 4 = 16$, som gir:

$$t_b = \frac{12}{\sqrt{(\frac{1}{2} * 9 * 8 - 0)(\frac{1}{2} * 9 * 8 - 16)}} = \frac{12}{\sqrt{720}} = 0.447 \tag{104}$$

Generelt har vi dersom en rangering inneholder en binær variabel med x og n-x = y medlemmer i hver av de to klassene (f.eks menn og kvinner) at: $\frac{1}{2}n(n-1) - V = \frac{1}{2}n(n-1) - \frac{1}{2}x(x-1) - \frac{1}{2}(n-x)(n-x-1) = xy$, som gir:

$$t_b = \frac{S}{\sqrt{(\frac{1}{2}n(n-1) - U)xy}} \tag{105}$$

, og for eksempelet over får vi:

$$t_b = \frac{12}{\sqrt{(\frac{1}{2} * 9 * 8 - 0)5 * 4}} = \frac{12}{\sqrt{720}} = 0.447 \tag{106}$$

4.3.4 Like rangeringer med to binære variable

I det ekstreme tilfellet får vi at begge rangeringene består av binære variable. Det kan da være fornuftig å fremstille informasjonen i en kontingenstabell jfr.

Tabell 11: Kontingenstabell for to binære variable 2

	Binær variabel 1		
Binær variabel 2	Ja	Nei	Totalt
Ja	a	b	k
Nei	c	d	l
Totalt	x	y	n

tabell 11.
Hvis du smln. et medlem, som faller i gruppen (ja,ja) (a stk) med et medlem i gruppen (nei,nei)(d stk) har vi et par med samme rekkefølge i begge rangeringene og hvert slikt par tilfører altså score 1 til summen S. Tilsvarende blir det om man kobler medlem fra (ja,nei)-klassen(c stk) med medlem fra (nei,ja)-klassen(b stk), men slik at hvert av disse parene tilfører verdien -1 til summen S. Det er åpenbart at andre koblinger av de fire klassene tilfører 0 til summen S. Dermed får vi jfr. tabell 11:

$$t_b = \frac{ad - bc}{\sqrt{xykl}} \tag{107}$$

Jeg bemerker at hvorvidt vi ser på de to binære variablene, som ordinale, eller nominale kan vi benytte t_b, som assosiasjonsmål. I tilfelle med to binære nominale variable ser vi dog bort fra fortegnet til t_b da retningen på assosiasjonen er uvesentlig.

4.4 Signifikanstester for r_s og Kendall's t når vi ønsker å teste om variablene er uavhengige - korrelasjon lik 0

r_s er en av mange mulige rangtester for uavhengighet($\rho_s = 0$) og var den første i rekken - antagelig grunnet sin enkelhet og likhet med Pearson's r. Dog vil alle fornuftige mål på korrelasjon mellom x og y basert på rang-verdier gi en test for uavhengighet. Daniels definerte, jfr. [16], en klasse korrelasjonskoeffisienter, som blant annet inkluderer r, r_s og Kendall's t og viste i [17] at disse målene essensielt bare er koeffisienter av uorden på den måten at hvis et par av verdier av y bytter plass for å bringe dem i samme rekkefølge, som korresponderende rekkefølge av x, vil enhver korrelasjonskoeffisient av nevnte klasse øke noe som for r_s sitt vedkommende klart fremkommer av uttrykket 85 på side 53.

Setter populasjonskorrelasjonskoeffisientene for r_s og Kendall's t til henholdsvis ρ_s og τ. Vi skal i denne seksjonen teste signifikansen til en observert korrelasjon Kendall's t eller r_s (H_0: Uavhengige variable(korrelasjon lik 0)). M.a.o. kan vi fra den observerte korrelasjonen konkludere med at det eksisterer en korrelasjon ρ_s eller τ ikke lik null for hele populasjonen?
Hvis det ikke er noen sammenheng mellom de to variablene X og Y, som vi ser på, vil det for et tilfeldig utvalg på n par (X,Y) være slik at en hvilken som helst rekkefølge for Y-variabelen vil være like sannsynlig for en gitt rekkefølge av X. Hvis vi velger en tilfeldig rekkefølge av X (det spiller ingen rolle hvilken så jeg velger den naturlige rekkefølgen 1,2,3,4,..,n) vil alle de n! mulige rekkefølgene av tallene 1 til n for Y være like sannsynlige - alle med sannsynlighet $\frac{1}{n!}$(dette er altså analogt med hvordan vi utførte permutasjonstester for Pearson's r i 3.1 på side 40). Vi vet videre at det til hver enkelt rangering av Y-variabelen vil korrespondere en verdi av Kendall's t eller r_s. Det totalt antall(n!) slike verdier kan vi klassifisere i henhold til den faktiske verdien av Kendall's t eller r_s i henhold til hva vi kan kalle en frekvensfordeling. Jeg skal benytte denne fordelingen til å illustrere hvordan vi kan utføre eksakte signifikanstester.
Vi studerer altså tester, som er basert på fordelingen av korrelasjonene i en populasjon, ved at vi permuterer observasjonene i alle mulige rekkefølger. Dette er en test av hypotesen om at variablene vi ser på er uavhengige i populasjonen. Testen vil avgjøre om en observert korrelasjon er signifikant forskjellig fra 0 i populasjonen eller ikke.
Vi skal i subseksjon 4.5 på side 70 se hvorvidt det lar seg gjøre å finne statistisk korrekte øvre og nedre grenser for en beregnet korrelasjon, som ikke nødvendigvis er null, relatert til et på forhånd satt signifikansnivå. Anta at vi får en korrelasjon på 0.5, som viser seg å være signifikant forskjellig fra null, kan vi da si mellom hvilke verdier den sanne korrelasjonen ligger? Og dersom vi i neste måling får en signifikant korrelasjon på 0.7 kan vi da si at denne er signifikant større enn den foregående på 0.5? Alle resulater, som benyttes i seksjonen her(4.4) og påløpende seksjon 4.5, er bevist i [1, s. 91-116 kap5].

4.4.1 Signifikanstester Kendall's t

Vi ønsker igjen å teste H_0: $\tau = 0$, der τ representerer Kendall's korrelasjonskoeffisient(t) fra populasjonen utvalget stammer fra.

$\mathbf{n < 10}$ Anta at vi har n=5 observasjoner og dermed 120 mulige måter å ordne de 5 observasjonene på. Hvis vi skriver ned alle mulige rekkefølger, og

Tabell 12: Frekvensfordelingen til S = P - Q

Verdien til S	Antall	Sannsynlighet
0	22	0.592
2	20	0.408
4	15	0.242
6	9	0.117
8	4	0.042
10	1	0.00083

beregner S for hver enkelt rekkefølge parret med den naturlige rekkefølgen 1,2,3,4,5 for den ene variabelen, finner vi frekvensfordelingen vist i tabell 12. I tabellen er det også opplyst om sannsynligheten for at S er større eller lik en spesifisert verdi S_0. De samme verdiene er sannsynligheten for at S er mindre eller lik den korresponderende negative verdien S_0.
Egenskaper ved frekvensfordelingen til S:

- Fordelingene er alltid symmetriske. Hvis $\frac{1}{2}n(n-1)$ er partall kan S bare ta partallsverdier og maksimumfrekvensen oppstår for S=0. Hvis $\frac{1}{2}n(n-1)$ er en oddetallsverdi kan S bare ta oddetallsverdier og vi får maksimumsfrekvenser for $\pm$ 1.

- Frekvensene er strengt synkende fra maksimum til frekvensen er 1 når S = $\pm\frac{1}{2}n(n-1)$

- Når n øker så tenderer formen på frekvenspolygonet mot normalkurven

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-x^2}{2\sigma^2}$$

  Når n>10 gir denne kurven en tilfredsstillende tilnærming til polygonet med med forventning 0 og $\sigma = \frac{1}{18}n(n-1)(2n+5)$

For n $\leq$ 30 fremkommer den eksakte fordelingen av [1, Appendix Table 1]. For n$\leq$10 viser tabellen sannsynligheten for at S er over eller lik en spesifisert positiv verdi S_0 og der den samme verdien er sannsynligheten for at S er mindre eller lik den korresponderende negative S_0. For 11$\leq$ n $\leq$ 30 er de kritiske verdiene for Kendall's t for utvalgte kvantiler(signifikansnivåer) oppgitt. I stedenfor å gå til denne, eller en annen tabell, bruker jeg statistikkprogrammet R til å beregne de aktuelle p-verdier. Vi ønsker først å teste hypotesen H_0: $\tau = 0$ og det er da naturlig å forkaste H_0 for høye verdier

av |t|(absoluttverdien til Kendall's t). Dersom vi på forhånd antar at det er enten positiv eller negativ korrelasjon ser vi bare på øvre eller nedre hale av fordelingen ved bestemmelse av signifikansnivå(ensidig test) - vi tester hypotesen H_{0_-}: $\tau < 0$ eller H_{0_+}: $\tau > 0$.
Generelt, ved bruk av signifikansstester, rapporters p-verdier. Vi definerer p-verdi = P(|S| $\geq$ |S_0|) når vi skal rapportere det tosidige alternativet H_A : $\tau \neq 0$ og p-verdi = P(S $\geq$ S_0) ,eller p-verdi = P(S $\leq$ S_0) når vi tester henholdsvis H_{A_+} : $\tau > 0$ eller H_{A_-} : $\tau < 0$. Dersom p-verdien er mindre enn det på forhånd satte signifikansnivået forkaster vi H_0, H_{0_+}eller H_{0_-}.

n > 10 Når n > 10 kan vi bruke normaltilnærmingen med:

$$E(S) = E(t) = 0 \tag{108}$$

$$Var(S) = \frac{1}{18}n(n-1)(2n+5) \tag{109}$$

$$\sigma_S = \frac{\sqrt{n(n-1)(2n+5)}}{3\sqrt{2}} \tag{110}$$

$$Var(t) = (\frac{2}{n(n-1)})^2\frac{1}{18}n(n-1)(2n+5) \tag{111}$$

$$\sigma_t = \frac{\sqrt{2(2n+5}}{3\sqrt{n(n-1)}} \tag{112}$$

Vi får at:

$$z = \frac{3S\sqrt{2}}{\sqrt{n(n-1)(2n+5)}} = \frac{3t\sqrt{n(n-1)}}{\sqrt{2(2n+5)}} \tag{113}$$

, som gir p-verdier for Kendall's t (og S) for henholdsvis H_0,H_{0_+} og H_{0_-} på:
$2P(z \leq \frac{3t\sqrt{n(n-1)}}{\sqrt{2(2n+5)}})$
$P(z \geq \frac{3t\sqrt{n(n-1)}}{\sqrt{2(2n+5)}})$
$P(z \leq -\frac{3t\sqrt{n(n-1)}}{\sqrt{2(2n+5)}})$

kontinuitetskorreksjon for S Normalfordelingen er kontinuerlig mens eksaktfordelingene til både Kendall's t og S er diskrete/diskontinuerlige. For å forbedre tilnærmingen av den eksakte fordelingen med normalfordelingen kan man gjøre en kontinuitetskorreksjon i observatoren z. Dette er ikke svært enkelt når z er en funksjon av Kendall's t, fordi forskjellene i mulige verdier

for Kendall's t ikke er konstant. Derimot er S-verdiene alltid partallsverdier eller alltid oddetallsverdier avhengig av om $\frac{1}{2}n(n-1)$ er partall eller oddetall. Verdiene til S skiller alltid på to enheter og vi kan derfor se på S,som uniformt fordelt, i intervallet S-1,S+1 istedenfor å være konsentrert helt og holdent på S. M.a.o minuserer vi 1 fra den observerte verdien på S hvis den er positiv og plusser 1 hvis den er negativ når vi beregner z for å finne aktuell p-verdi.

Like rangeringer(ties) Akkurat som når vi ikke har likt rangerte data vil fordelingen til Kendall's t for et fiksert antall likt rangerte data tendere mot normalitet når n øker og det er antagelig lite feil å gjøre ved å bruke normaltilnærmingen for n $>$ 10 med mindre det er en stor gruppe med samme rangering eller mange grupper med like rangeringer.

To variable med henholdsvis u og v likt rangerte data Dersom det er u likt rangerte observasjoner i den ene variabelen X og v likt rangerte observasjoner i den andre variabelen Y vil variansen til fordelingen til S, som oppdages ved å korrelere en rangering med alle n! mulige rangeringer av de andre, være gitt ved:

$$var(S) = \frac{1}{18}(n(n-1)(2n+5) - \sum u(u-1)(2u+5) - \sum v(v-1)(2v+5)) \tag{114}$$

$$+\frac{1}{9n(n-1)(n-2)}(\sum u(u-1)(u-2))(\sum v(v-1)(v-2)) \tag{115}$$

$$+\frac{1}{2n(n-1)}(\sum u(u-1))(\sum v(v-1)) \tag{116}$$

Kun en variabel inneholder likt rangerte observasjoner Dersom bare en rangering inneholder like rangeringer slik at f.eks. alle v=0 har vi:

$$var(S) = \frac{1}{18}(n(n-1)(2n+5) - \sum u(u-1)(2u+5)) \tag{117}$$

En binær variabel og en variabel med likt rangerte observasjoner Hvis en variabel degenereres til en binær variabel med x og (n-x)=y medlemmer og likt rangerte data i den andre variabelen defineres ved u får vi:

$$var(S) = \frac{xy}{3n(n-1)}(n^3 - n - \sum(u^3 - u)) \tag{118}$$

To binære variable Blir begge variablene binære får vi:

$$var(S) = \frac{xykl}{n-1} \tag{119}$$

Variansuttrykkene ovenfor gir oss muligheten for å teste signifikansen til t_a, t_b, t_c, eller identisk signifikansen til verdiene til S, som t_a, t_b og t_c er utledet fra. Vi får dog et problem knyttet til kontinuitetsbetraktninger:

- I tilfelle med en binær variabel og en variabel uten likt rangerte data er intervallet mellom sukkessive verdier av S lik 2. Den riktige endringen på S er halvparten av 2.
- I tilfelle med en binær variabel og en rangering med bare u likt rangerte observasjoner blir intervallet til S 2u og riktig endring i S ved kontinuitetskorreksjon er u.
- Når begge variablene er binære er intervallet n mellom mulige verdier av S og den riktige endringen på S ved kontinuitetskorreksjon blir n/2.
- Hvis en variabel er dikotomisert og den andre inneholder likt rangerte data med ulikt antall vil det være varierende forskjeller i avstandene mellom mulige verdier av S og i så fall må vi bruke en passende approksimativ metode ved beregning av kontinuitetskorreksjonen.

4.4.2 Signifikanstester r_s

Vi ønsker igjen å teste H_0: $\rho_s = 0$, der ρ_s representerer Spearman's rank-korrelasjonskoeffisienten fra populasjonen utvalget stammer fra.

n≤35 Akkurat som for Kendall's t er fordelingen til r_s symmetrisk om 0 og fordelingen tenderer mot normalitet for store n, men langt saktere enn hva Kendall's t gjør. Fordelingen til r_s for en gitt n er fullstendig definert ved fordelingen til $\sum d^2$.
Anta at vi har n=5 observasjoner og dermed 120 mulige måter å ordne de 5 observasjonene på. Hvis vi skriver ned alle mulige rekkefølger og beregner $\sum d^2$ for hver enkelt rekkefølge parret med den naturlige rekkefølgen 1,2,3,4,5 finner vi sannsynlighetsfordelingen vist i tabell 2 i appendikset til [1]. Tabellen viser for n ≤ 16 bare de verdier hvor $\sum d^2 > n(n^2-1)/6$ ($r_s < 0$), som er de med venstrehale p-verdier ≤ 0.5. Ved symmetri kan vi finne høyrehale p-verdiene for $\sum d^2 < n(n^2-1)/6 (r_s > 0)$ ved å gå inn i tabellen for verdien $n(n^2-1)/3 - \sum d^2$. For eksempel hvis n=5 og $\sum d^2 = 12$ kan du gå inn i tabellen for 40 - 12 = 28 og lese av den høyre p-verdien på 0.258. For 17

$\leq$ n $\leq$ 35 gir tabellen verdier for r_s for utvalgte kvantiler - korresponderende signifikanspunkter for negative r_s er de samme ved symmetri. I stedenfor å gå til denne, eller en annen tabell, bruker jeg statistikkprogrammet R til å beregne de aktuelle p-verdier.

n > 35 For n>35 er det typisk godt nok å bruke normaltilnærmingen med:

$$E(r_s) = 0 \tag{120}$$

$$var(\sum d^2) = (\frac{n^3 - n}{6})^2 \frac{1}{n-1} = \frac{n^2(n-1)(n+1)^2}{36} \tag{121}$$

$$\Updownarrow \tag{122}$$

$$var(r_s) = \frac{1}{n-1} \tag{123}$$

$$\sigma_{r_s} = \frac{1}{n-1} \tag{124}$$

Vi får at:

$$z = r_s\sqrt{n-1} = \sqrt{n-1}(1 - \frac{6\sum d^2}{n^3 - n}) \tag{125}$$

, som korresponderer til p-verdier fra sannsynlighetsfunksjonen til standard-normalfordelingen. Dersom p-verdien for en gitt observator er lavere enn et på forhånd fastsatt signifikansnivå forkastes nullhypotesen.

Kontinuitetskorreksjon for $\sum d^2$ $\sum d^2$ kan bare ta partallsverdier mellom 0 og $\frac{n^3-n}{3}$ med forventning $\frac{n^3-n}{6}$. For å gjøre kontinuitetskorreksjonen trekker vi derfor fra 1 på den observerte verdien til $\sum d^2$ hvis den er mindre enn forventningen på $\frac{n^3-n}{6}$ og legger til 1 dersom $\sum d^2$ overstiger $\frac{n^3-n}{6}$ når vi skal beregne z-verdien for å finne aktuell p-verdi. Bemerker at det er komplisert å inkorperere kontinuitetskorreksjon i uttrykkene som inneholder r_s, fordi forskjellene i mulige verdier til r_s ikke er konstante. Dette taler i favør av å bruke de uttrykk, som bare inneholder d^2, når det er ønskelig å bruke kontinuitetskorreksjon.

Like rangeringer Hvis vi har observasjoner, som for en eller begge variablene, er likt rangert påvirker ikke dette variansen til r_s. Vi får dog problemer med hensyn på kontinuitetskorreksjon ved likt rangerte data.

- For en binær variabel mot en variabel uten likt rangerte observasjoner er avstanden mellom verdiene til $\sum d^2$ lik n og vi endrer med $\pm \frac{n}{2}$ på $\sum d^2$.

- For en binær variabel mot en variabel som kun består av et visst antall likt rangerte observasjoner er intervallet n*u mellom verdiene til $\sum d^2$ og og kontinuitetskorreksjonen blir $\pm \frac{n*u}{2}$
- Dersom vi har dobbelt dikotomi blir kontinuitetskorreksjonen $\pm \frac{n^2}{4}$

Det er verdt å bemerke at dersom vi har dobbel dikotomi vil r_{sb} være lik t_b og vi bruker da typisk den siste ved beregninger.

Bruk av målefordelingen til Pearson's r Det følger intuitivt og er lett å vise at r_s kan bli sett på som Pearson's r mellom rangeringer m.a.o. slik at rangeringene anses, som de observerte kontinuerlige variable - for bevis se [1, s. 27].
Siden formlene for momentene til den generelle målefordelingen til permutasjonsfordelingen til r holder for alle x og y holder de også for r_s. Videre er oddetallsmomentene til de naturlige tallene om gjennomsnittet lik null, som følge av symmetri. Vi får følgende momenter, jfr. [3, s. 495], for r_s:

$$E(r_s) = 0 \tag{126}$$

$$var(r_s) = E(r_s^2) = \frac{1}{n-1} \tag{127}$$

$$E(r_s^3) = 0 \tag{128}$$

$$E(r_s^4) = \frac{3}{n^2-1}(1 + \frac{12(n-2)(n-3)}{25n(n-1)^2}) \tag{129}$$

For n $\geq$ 10 er approksimasjonen:

$$dF = \frac{1}{B(\frac{1}{2}, \frac{1}{2}(n-2))}(1-r^2)^{\frac{1}{2}(n-4)}dr, -1 \leq r \leq 1 \tag{130}$$

adekvat for nær alle praktiske formål da også $E(r_s^4)$ er neglisjerbar for n$\geq$ 10 - se tabell 13 på neste side over eksakte og approksimerte kritiske verdier for r_s for n=10. Dette siste resultatet følger av resultatet 83 på side 45 til den generelle målefordelingen til r.

4.5 Signifikanstester og konfidensintervaller

Testene vi har sett på til nå er basert på fordelingen til korrelasjonene oppdaget ved å permutere rangeringene i alle mulige rekkefølger. Vi har testet hvorvidt de to variablene vi studererer, relatert til deres rangeringer,

Tabell 13: Eksakte og approksimerte kritiske verdier for Spearman's r_s

Smln. av eksakte og approksimerte kritiske verdier til r_s for n=10		
Tosidig test	Eksakt(Kendall(1955))	Approksimert
$\alpha = 0.05$	0.648	0.632
$\alpha = 0.01$	0.794	0.765

er uavhengige i populasjonsfordelingen. Testene vi har sett på viser altså om en observert korrelasjon er signifikant forskjellig fra 0.
Hvis vi så har funnet en utvalgskorrelasjon signifikant forskjellig fra null kan vi da si mellom hvilke verdier den sanne populasjonskorrelasjonen ligger? Hvis vi videre i neste måling får en annen korrelasjon enn i den første kan vi da hevde at denne er ulik den første? Jeg vil først illustrere problemstillingene, som oppstår når vi prøver å finne svaret på de to nevnte spørsmål, ved å anta at populasjonsfordelingen er endelig lik en gitt N før jeg ser på mulighetene for å lage konfidensintervaller jfr. [1, s. 72-76].
La oss anta at hele populasjonen(N), som utvalget(n) stammer fra, er rangert i henhold til den ene variabelen X i rekkefølgen 1,2,...,N. Lar så rangeringen for den andre variabelen være p_i for det i'te medlemmet. Vi kan så beregne populasjonskorrelasjonskoeffisienten på vanlig måte, men nå for hele populasjonen. Anta så at vi, som vanlig, gjør et utvalg på n av de N. Disse n medlemmene vil da selvsagt være i naturlig rekkefølge i forhold til variabelen X og vi kan beregne Kendall's t på vanlig måte. Til hvert mulige utvalg av n vil det korrespondere en verdi av Kendall's t og siden det er $\binom{N}{n} = \frac{N!}{n!(N-n)!}$ slike utvalg vil det være et identisk antall verdier av Kendall's t korresponderende til alle mulige utvalg. Vi kan vise, jfr. [1, s. 95-97], at uansett populasjonsfordeling vil fordelingen til Kendall's t beregnet for alle mulige utvalg tendere mot normalfordelingen når n øker forutsatt at τ ikke er for nær yttergrensene -1 og 1 og visse ikke spesielt begrensede betingelser er oppfylt. Man kan også vise at forventningsverdien til utvalgsfordelingen(og dermed for den asymptotiske normalfordelingen) er τ, men derimot oppstår det problemer ved beregning av standardavviket til utvalgsfordelingen da standardavviket avhenger av andre størrelser enn τ og standardavviket vil variere med rangeringen til den andre variabelen i motsetning til forventningsverdien til Kendall's t.
Kendall og Gibbons(1990) illustrerer dette på en enkel måte. Anta at den første variabelen(X) gitt populasjonen N=9 er rangert i naturlig

Tabell 14: Fordelingen til P(overensstemmende par) gitt N=9 og n=3 for rangeringene A og B

Fordelingen til P gitt N=9 og n=3 for rangering A og rangering B		
Verdiene til P	Frekvens rangering A	Frekvens rangering B
0	2	3
1	15	16
2	34	29
3	33	36
Total	84	84

rekkkefølge(1,2,3,...,9) og at rangeringen av den andre variabelen(Y) da er 5,2,3,1,6,7,8,9,4(rangering A). Lar vi det tilfeldige utvalget være på n=3 får vi $\binom{9}{3} = 84$ mulige utvalg. Hvis vi beregner P(antall overensstemmende par) for hvert av de 84 utvalgene får vi fordelingen jfr. tabell 14 kolonne 2. Forventningsverdien til denne fordelingen finner vi kjapt at er 13/6 og dermed at $E(t) = \frac{4P}{n(n-1)} - 1 = \frac{4*\frac{13}{6}}{3(3-1)} - 1 = 0.44$. Vi finner lett, for populasjonskorrelasjonskoeffisienten τ relatert til den endelige populasjonen med N=9, at P=26 og dermed at $\tau = \frac{4P}{n(n-1)} - 1 = \frac{4*26}{9(9-1)} - 1 = 0.44 = E(t)$, som bekrefter vår antagelse om at forventningsverdien til t lik populasjonskorrelasjonskoeffisienten τ.
La oss så se på rangeringen 1,2,5,9,3,6,7,8,4(rangering B) av den andre variabelen(Y). Vi får igjen at τ=0.44, men fordelingen til P ved et utvalg på n=3 blir nå som vist i 14 kolonne 3. Vi får også nå at E(t)=τ=0.44, men fordelingen til P under rangering B er uansett forskjellig fra fordelingen til P under rangering A. Vi får blant annet at variansene for de to fordelingene blir 0.734 og 0.639, for henholdsvis rangering A og B, og altså ulike.
Vi får altså problemer med å uttrykk var(t) ved hjelp av kjente faktorer med mindre vi vet noe om rangeringene i populasjonen(sjeldent tilfelle).
Man kan dog vise at for en hvilken som helst populasjonsfordeling vil variansen til Kendall's t ikke overstige:

$$var(t) \leq \frac{2}{n}(1 - \tau^2) \tag{131}$$

Dette resultatet holder også bra når vi har likt rangerte observasjoner og ved å bruke denne variansen kan vi utføre en test på den sikre siden når vi kan anta at fordelingen til Kendall's t er normalfordelt og derigjennom kan bruke

den estimerte Kendall's t , som forventningsverdi.
Et tilsvarende approksimativt uttrykk for r_s for **store n** er:

$$var(r_s) \leq \frac{3}{n}(1-\rho_s^2) \tag{132}$$

, men uttrykket er ikke verifisert for å gjelde når vi har likt rangerte observasjoner.
For resultater og bevis se [1, s. 104-116]

4.5.1 H_0: $\tau = \tau_q$ og H_0: $\rho_s = \rho_q$

1) H_0: $\tau = \tau_q$

Anta at vi har n=30 og at vi har funnet at Kendall's t=0.59 for dette tilfeldige utvalget. I tilfelle med n=30 kan fordelingen til Kendall's t anses normalfordelt. Hva kan vi da slutte om verdien av populasjonskorrelasjonen τ i populasjonsfordelingen utvalget er tatt fra? Dersom Kendall's t = 0.59 og n = 30 får vi at var $t \leq \frac{2}{30}(1-0.59^2) = 0.04346$, som gir et standardavvik på 0.208. Sannsynligheten for at et avvik fra gjennomsnittet er 1.96 standardavvik eller mer (i absoluttverdi) er 0.05. Vi kan derfor si at på det verste vil det være 0.95 sjanse for at den sanne verdien til τ ligger i intervallet 0.59 ± 1.96*0.208 = (0.18,0.998), m.a.o. at sjansen er mindre eller lik 0.05 for at den virkelige verdien til τ ligger utenfor intervallet. Istedenfor å definere øvre og nedre grense i intervallet på bakgrunn av en ønsket sannsynlighet(signifikansnivå α) har vi her satt grensene m.h.p en maksimal grad av sannsynlighet, eller sagt på en annen måte setter vi yttergrensene til populasjonsfordelingens τ og konfidensintervallene kan sies å være konservative. I eksempelet her har vi erstattet den ukjente populasjonskorrelasjonskoeffisienten τ med Kendall's t, men vi kan omgå dette ved å benytte oss av teorien knyttet til konfidensintervaller. Når vi antar at Kendall's t er normalfordelt N(t,$\frac{2}{n}(1-t^2)$), og x er antall standardavvik i absoluttverdi fra forventningsverdien relatert til et gitt signifikansnivå α under standardnormalfordelingen(N(0,1)), vil $\tau_1 \leq \tau \leq \tau_2$, der τ_1 og τ_2 er røttene til: $t-\tau = x\sqrt{\frac{2}{n}(1-\tau^2)}$, som gir $\tau = \frac{t \pm x\sqrt{2/n}\sqrt{1+2x^2/n-t^2}}{1+2x^2/n}$. Denne mer vanlige måten å konstruere konfidensintervaller for τ på er en mer nøyaktig måte enn den første vi så på og slik at vi kontrollerer 100% for type-1 feil. Med eksempelet ovenfor får vi konfidensintervallet (0.09,0.85).

2) H_0**:** $\rho_s = \rho_q$

Vi får tilsvarende resultater, som for τ ved bruk av:

$$var(r_s) \leq \frac{3}{n}(1 - \rho_s^2) \tag{133}$$

, men slik at normalfordelingen tilnærmes langt saktere og slik at n typisk må være høy og det er foreløpig heller ikke verifisert at dette resultatet for variansen til r_s gjelder ved likt rangerte data. Jeg bemerker at siden r_s bare er et spesialtilfelle av r kan alle resultater relatert til tilfeller der populasjonsfordelingen er normalfordelt benyttes jfr. 2.5.2 på side 21.

4.5.2 H_0**:** $\tau_q = \tau_p$ **og** H_0**:** $\rho_q = \rho_p$

H_0**:** $\tau_q = \tau_p$ Anta at vi i et tilfeldig utvalg fra populasjonen har n_1=20 med observator $t_1 = 0.8$ og i et annet tilfeldig utvalg fra samme populasjon har n_2=20 og observator $t_2 = 0.6$. Det er da naturlig å spørre om hvorvidt $\tau_1 = \tau_2$. En angrepsmetode er å se om konfidensintervallene for τ_1 overlapper konfidensintervallet for τ_2 eller motsatt. Relatert til t_1 har vi $var(t) = \frac{2}{n}(1 - t^2) = 0.036$, som gir et standardavvik på 0.19 og siden forskjellen på t_1 og t_2 bare er 0.20 kan vi umiddelbart konkludere med at forskjellen ikke er signifikant.
Generelt er det slik at hvis vi har et antall verdier av S(selv fra rangeringer med ulik størrelse) kan vi plusse dem sammen og teste signifikansen til hele settet med summen av variansene til de individuelle rangeringene. Poenget er at variansen til en sum av uavhengige variable er summen av variansene. I tilfellet med de to målingene ovenfor er maksimumsvariansen til t_2 på 0.064 slik at et maksimum til variansen til forskjellene mellom de to estimatene for τ er 0.1 med korresponderende standardfeil på 0.32 - standardfeilen er større enn den faktiske forskjellen på 0.2 og igjen må vi konkludere med at forskjellen ikke er signifikant.

H_0**:** $\rho_q = \rho_p$ Vi får tilsvarende resultater, som for τ ved bruk av:

$$var(r_s) \leq \frac{3}{n}(1 - \rho_s^2) \tag{134}$$

, men slik at normalfordelingen tilnærmes langt saktere og slik at n typisk må være høy og det er foreløpig heller ikke verifisert at dette resultatet for variansen til r_s gjelder ved likt rangerte data.

4.5.3 Inferensproblemer

Inferens knyttet til korrelasjonskoeffisientene Kendall's t og r_s er som vi har sett forbundet med store standardfeil. Uansett hva τ eller ρ_s er vil standardfeilene til Kendall's t og r_s være henholdsvis av orden $\sqrt{2/n}$ og $\sqrt{3/n}$. Dette er dog generelt en dårlig egenskap ved korrelasjonskoeffisienter - tidligere har vi sett at standardfeilen til Pearson's r ved store målinger funnet ved trekk fra bivariate normalfordelinger er $(1-\rho^2)/\sqrt{n}$ og er av orden $1/\sqrt{n}$. Det er altså generelt vanskelig å lokalisere populasjonskorrelasjonen veldig nøyaktig med mindre n ligger i intervallet 30-40 eller høyere. Poenget er at man skal være svært forsiktig med og legge for mye i korrelasjonskoeffisienter beregnet fra data hvor n er lav med mindre vi har mange tilgjengelige målinger.
For eksempel med n=32 blir maksimal standardfeil $\frac{1}{4}\sqrt{1-t^2}$ og hvis Kendall's t er nær 0 vil vi ikke med et 95% konfidensintervall kunne lokalisere populasjonskorrelasjonskoeffisienten bedre enn at den må befinne seg ca. i intervallet (-0.5,0.5). Hvis Kendall's t=0.8 blir lokaliseringen noe bedre, men fortsatt vil Kendall's t kunne ligge i intervallet (0.5,1).

4.6 Eksistens og avhengighet av fordelingsantagelser

Gyldigheten og bruken av metodene diskutert i seksjonen her avhenger ikke på noen måte av underliggende fordeling på X,Y og det lar seg alltid gjøre å beregne Kendall's t og r_s så lenge vi klarer å rangere data.

4.7 Robusthet

I kontrast til Pearsons r krever ikke ordinære ikke-parametriske korrelasjonsmål avhengighet av en bivariat normalfordeling og kan, som nevnt tidligere, videre brukes på både parrede observasjoner av kontinuerlige data og data bestående av rangeringer. Rangeringene kan enten være de originale data, eller de kan utledes fra kontinuerlige målinger.
Ikke-parametriske korrelasjonsmål behandlet i seksjonen her er også robust mot uteliggere.

4.8 Utvalgsstørrelse og inferens

Testene for å teste uavhengighet kan brukes for $n \geq 4$, men som vi har sett krever vi at fordelingen til Kendall's t og r_s må være tilnærmet normal før vi kan beregne konfidensintervaller knyttet til et beregnet estimat, eller for å se om det eksisterer forskjeller på to beregnede estimater.

Vi har videre sett at det er svært vanskelig å lokalisere den virkelige populasjonskorrelasjonen med mindre n er høy, eller vi har mange målinger, jfr. de høye maksimumsvariansene $var(t) = \frac{2}{n}(1-\tau^2)$ og $var(r_s) = \frac{3}{n}(1-r_s^2)$. Man må derfor utvise stor grad av forsiktighet før det blir aktuelt å gjøre inferens og n bør under enhver omstendighet være på omlag 30-40. I oppgaven her vil jeg studere datasett med n=5 og n=6 slik at inferens knyttet til bruk av Kendall's t og r_s, utover å teste $H_0 : \rho = 0$, blir uaktuelt.

4.9 Nærmere om valget mellom r, r_s og Kendall's t

Vi har sett at r,r_s og Kendall's t er spesialtilfeller av den generelle korrelasjonskoeffisienten Γ. Γ viser også hvordan vi får ulike koeffisienter i henhold til vår måte å score forskjellene på mellom rangeringene av to målinger. Scoringsmetoden for Kendall's t er den enklest mulige der hver scoring har verdien 1 uavhengig av hvor stor avstanden er mellom rangeringene på de to målingene. Scoringsmetoden for r_s gir større vekt til forskjeller mellom rangeringene jo større avstanden er mellom rangeringene. Scoringsmetoden for r gir en objektiv verdi til forskjellen mellom verdiene på rangeringene på den måten at forskjellene måles direkte på den kontinuerlige skalaen, som x og y ble observert på i utgangspunktet. Hvilken av metodene, eller eventuelt andre alternative metoder, man skal bruke avhenger av hvilke metoder som er mulige å bruke m.h.p. målestørrelsen n og praktiske overveielser.
I enkelte tilfeller der størrelser varierer stort, kan bruk av originale kontinuerlige data medføre en lite retningsgivende relasjon mellom to variable ved at en eller to ekstremt store tall mer eller mindre eliminerer effekten av de små. I slike tilfeller kan Pearson's r være et dårlig alternativ og det vil være bedre å erstatte variabelverdiene målt på en kontinuerlig eller forholdstallsskala med rangeringer for å gjenopprette en balanse.
Hvis utvalget er under 11 og vi ikke har data fra en bivariat normalfordeling, er det naturlig å velge r_s, Kendall's t, evt. en generell permutasjonsfordeling for Pearson's r, eller, hvis man tror målefordelingen er representativ for en antatt tilnærmet fordeling $\hat{F}$ for F, bootstrapping.
Når vi har med ordinale data å gjøre opererer vi med rangerte data og valget står typisk mellom r_s og Kendall's t. I tilfeller hvor det bare er av viktighet å vite om antall uoverensstemmende par, blant de to rangerte variablene, er det naturlig å bruke Kendall's t. Hvis avstanden mellom posisjonene mellom de to rangeringene, som er med i et par, er av betydning i tillegg til om posisjonene er inverse(uoverensstemmende) eller ikke, så er det naturlig å bruke r_s.
Generelt har vi følgende sammenheng mellom r_s og Kendall's t:
Med bakgrunn i formlene $t = 1 - \frac{4Q}{n(n-1)}$ og $r_s = 1 - \frac{6\sum d^2}{n^3-n}$ har vi ved enkel

manipulasjon at $r_s = t$ når $\sum d^2 = \frac{2}{3}Q(n+1)$ og forholdet $\frac{\sum d^2}{Q}$ er da for en gitt n lik $\frac{2}{3}(n+1)$. Vi har videre at hvis $\sum \frac{d^2}{Q} > \frac{2}{3} * (n+1)$ er $r_s < t$ og hvis $\sum \frac{d^2}{Q} < \frac{2}{3} * (n+1)$ er $r_s > t$.
Kendall's tau er sammenlignet med r_s, som vi har sett tidligere, mer intuitiv å forstå og gir et bedre estimat(herunder mer eksakte p-verdier jfr. 4.5.3 på side 75 om ordenen på standardfeilen til Kendall's t og r_s på henholdsvis $\sqrt{2/n}$ og $\sqrt{3/n}$) for den korresponderende populasjonsparameteren spesielt for små målinger.
Under hypotesen om uavhengighet $H_0 : \rho = 0$ er Kendall's t og r_s høyt korrelerte og Daniels(1944) viste at denne korrelasjonskoeffisienten er gitt ved $\frac{2(n+1)}{(2n(2n+5))^{\frac{1}{2}}}$. Den nevnte korrelasjonskoeffisienten synker fra 1 når n=2 til 0.98 når n=5 for deretter å gå mot 1 når n går mot ∞. Testene er altså asymptotisk ekvivalente når H_0 holder. Daniels(1944) viste at den asymptotiske simultanfordelingen til r_s og Kendall's t når H_0 holder er bivariat normal. Når man trekker fra en bivariat normalfordeling og populasjonskorrelasjonskoeffisienten $\rho \neq 0$ er det fortsatt høy korrelasjon mellom r_s og Kendall's t jfr. [18], som viste at når n går mot inf, vil korrelasjonen mellom Kendall's t og r_s være $\geq$0.984 hvis $|\rho| \leq 0.8$ og en verdi ≥ 0.937 når ρ =0.9.
For å være på den sikre siden i forhold til å opplyse om for høye korrelasjoner kan det generelt være fornuftig å velge i henhold til min(|r_s|,|Kendall's t|)

4.10 Forholdet mellom varianser(effisiens) og uavhengighetstester

Jakten på distrubisjonsfrie statistiske tester er motivert av ønske om å utvide rekkevidden av gyldigheten av å kunne utføre inferens på 'mønsterløse data'. Ulempene ved en slik generalisering er at man selvsagt taper effisiens i enkelte sammenhenger - vi kan ikke forvente at en fordelingsfri test, som er valgt uavhengig av formen på populasjonsfordelingen, er like effisient som den testen vi ville valgt dersom vi hadde kjent populasjonsfordelingen. Det er likevel feil å bruke dette som et argument mot fordelingsfrie prosedyrer da det nettopp er fraværet av informasjon relatert til en populasjonsfordeling, som medfører at vi velger en fordelingsfri test. Dersom vi må ty til bruk av fordelingsfrie tester bør vi velge den mest effisiente av disse testene. Det er vanskelig å finne fordelingsfrie tester, som har høyest styrke sammenlignet med alle mulige alternativer og vi velger derfor å se på styrken av fordelingsfrie tester mot parametriske alternativer. Undersøker vi styrke mot alternativene relatert til normalfordelingsteori får vi et mål på hvor mye vi kan tape ved bruke en

fordelingsfri test hvis normalantagelsen faktisk hadde vært gyldig. Hvis dette tapet er lite er det en motivasjon for å ofre et lite tap av effisiens smln med å få en utvidet rekkevidde på gyldigheten knyttet til bruksområde for den fordelingsfrie testen. Jeg understreker at det dog ikke er noen grunn til å forvente at metoder som bygger på normalteori opprettholder sine effisiensfordeler i forhold til fordelingsfrie tester når populasjonsfordelingen ikke faktisk er normal.
Definerer de asymptotiske relative effisiensene(ARE) for permutasjonstesten til r, Kendall's t-testen og r_s-testen for uavhengighet relativt til den ordinære målekorrelasjonskoeffisienten r når den alternative hypotesen er at X,Y kommer fra en bivariat normalfordelig der $\rho \neq 0$:
$ARE_{perm,r} = 1$ jfr. [3, p. 499]
$ARE_{t,r} = \frac{9}{\pi^2} = 0.91$ jfr. [3, p. 499]
$ARE_{r_s,r} = \frac{9}{\pi^2} = 0.91$ jfr. [3, p. 499]
Bortsett fra effisiensresultatene relatert til at X,Y kommer fra en bivariat normalfordeling er det utført lite arbeid på effisienstester knyttet til uavhengighet spesielt fordi det er vanskelig å spesifisere ikke-normale alternativer når man skal teste uavhengighet. Dog har Konijn jfr. [19] studert en klasse av alternative fordelinger på X,Y(generert ved lineærtransformasjoner på to uavhengige variable). Konijn fant ut at Kendall's t og r_s ofte er asymptotisk ekvivalente tester (akkurat som når X,Y kommer fra en bivariat normalfordeling) og at begge tester har en ARE svært nære testen som er basert på målekorrelasjonskoeffisienten r. Med andre ord er i disse tilfeller $ARE_{perm,r}$, $ARE_{t,r}$ og $ARE_{r_s,r}$ svært like.

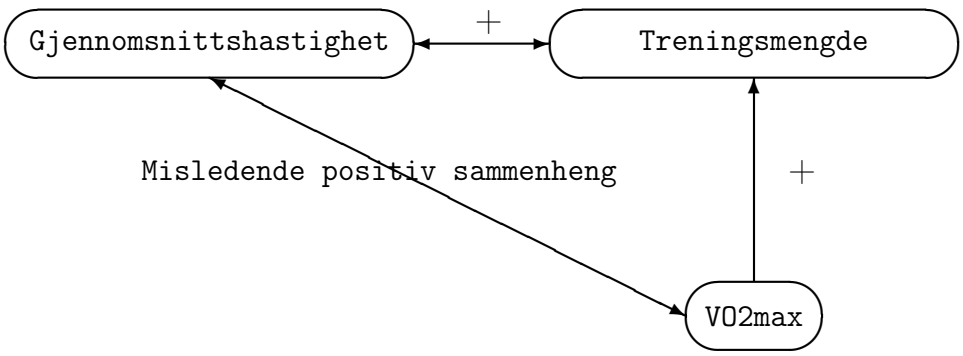
5 Misvisende korrelasjon

Selvom to størrelser X og Y korrelerer er det farlig å trekke konklusjoner kun på bakgrunn av korrelasjonen vi finner. Vi må: 1) være sikre på at vi har brukt riktig korrelasjonsmål for måledataene, 2) kunne si noe om korrelasjonen er statistisk signifikant, og 3) vi må diskutere hvorvidt korrelasjonen kan være misledende.
Ofte er to variable statistisk relatert, men det eksisterer likevel ingen årsakssammenheng mellom dem. Generelt vil korrelasjon støtte forklaringen om årsakssammenheng, men beviser den ikke!
En beregnet signifikant korrelasjon kan fint oppstå ved tilfeldigheter, eller fordi man har oversett en, eller flere andre variable, som påvirker en eller begge variablene X og Y. Det er m.a.o. svært viktig å ha klart for seg om det foreligger en kausal sammenheng mellom variablene X og Y, og om de

nevnte variablene igjen står i kausale forbindelser med andre variable. Først når man har kontroll på de kausale forbindelser vil det være mulig å interpretere beregnede korrelasjoner på en god måte, og derigjennom hindre at korrelasjonen man oppgir i for stor grad er misvisende, eller illusorisk.

5.1 Tredje variabler - partiell korrelasjon

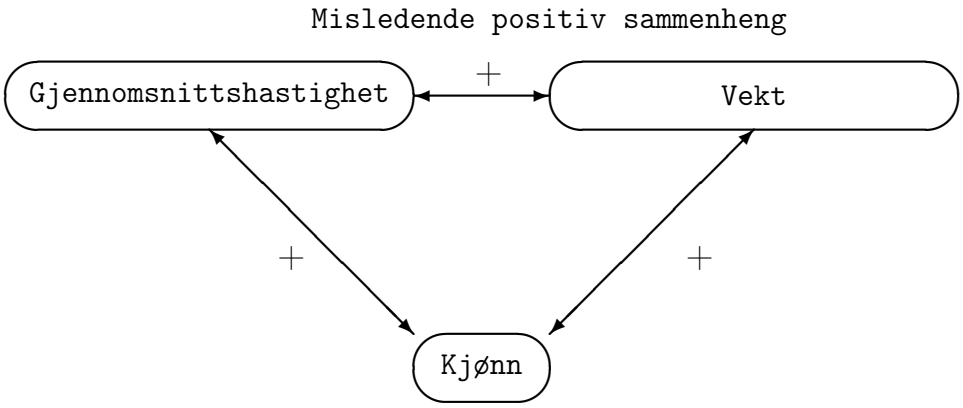
Det kan være en tredje variabel, Z, som påvirker både X og Y på en slik måte at det oppstår korrelasjon mellom størrelsene X og Y og korrelasjonen mellom X og Y er da misledende.
Dersom vi antar at det eksisterer en misledende sammenheng mellom to variable, som følge av påvirkning fra en eller flere andre variable kan vi teste dette ved å bruke en statistisk kontrollprosedyre, som partiell korrelasjon [3, 330-332], eller alternative angrepsvinkler som kovariansanalyse [3, 332-334] og hierarkisk regresjonsanalyse [3, 334-335]. Hvis vi ønsker å kontrollere kun for en enkelt tredje variabel kan vi benytte oss av en førsteordens partiellkorrelasjon. Dersom det eksisterer to eller flere variable vi ønsker å kontrollere for kan vi benytte oss av høyere ordens partiellkorrelasjoner. En førsteordens partiellkorrelasjon referer til en sammenheng mellom variablene X og Y etter at effekten av den tredje variabelen Z er fjernet (statistisk kontrollert for). At den partielle korrelasjonen er lik null er ikke tilstrekkelig til å si at den opprinnelige beregnede korrelasjonen(ikke lik null) er illusorisk. De Kausale forbindelser mellom variablene X,Y og Z er av stor betydning når man skal interpretere partielle korrelasjoner jfr. [34], som gir detaljert innsikt i temaet. Under antagelsen om at Z er den eneste variabelen som kan påvirke variablene X og Y, og at Z ikke påvirkes av verken X eller Y er den førsteordens partielle korrelasjonen den genuine korrelasjonen [34] og relasjonen mellom X og Y, uten å kontrollere for Z, anses å være illusorisk.
Ovenfor har jeg tydeliggjort at en relasjon gjennom korrelasjon mellom to variable ikke i seg selv impliserer årsakssammenheng mellom variablene. I denne sammenheng er det verd å bemerke at forskjellen på uavhengige og avhengige variable ikke er relevant i en korrelasjonskontekst(Hays, 1994).
VO2max korrelerer sterkt med prestasjonen i utholdenhetsidretter jfr. [51] og [50]. Selvom også vi, på f.eks. 10km klassisk kvinner på Beitostølen November 2014, beregnet en statistisk signifikant positiv korrelasjon mellom VO2max og gjennomsnittshastigheten på hele løpet, kan vi ikke ubetinget hevde at høyere VO2max vil gi(er årsaken til) høyere gjennomsnittshastighet på 10 km klassisk for Norges beste kvinner i langrenn. Det kan hende at denne sammenhengen er misledende grunnet de nevnte variablenes relasjon til en tredje variabel, som f.eks treningsmengde. Anta at forskning viser at



Figur 3: Misledende korrelasjon 1

individer med høy VO2max trener mer - da vil en høyere VO2maks påvirke treningsmengde positivt, som vist i figur 3. Det er også ofte blitt bekreftet at individer med høy treningsmengde inntil en viss mengde presterer bedre, og uten tap av generalitet har jeg, som vist i figur 3, ikke lagt noen antagelse på årsakssammenhengen mellom treningsmengde og gjennomsnittshastighet på hele løpet (jfr. linjen med pil på begge sider markert med + for positiv sammenheng). Konklusjonen i henhold til relasjonsmønstrene gitt i figur 3 er at man kan observere en positiv korrelasjon mellom VO2max og gjennomsnittshastighet på hele løpet, selvom den virkelige relasjonen er null.

En misledende sammenheng kan også observerers når en tredje variabel simultant kan relateres til to andre urelaterte variable(Simpsons paradoks). F.eks kan man observere en positiv sammenheng mellom kroppsvekt og gjennomsnittshastighet på hele løpet når man studerer kvinnelige og mannlige langrennsløpere i norgestoppen under et. Jeg kjenner ikke til at det er påvist at vekt på Norges toppløpere, for kvinner eller menn, skal være positivt korrelert med gjennomsnittshastighet på hele løpet(jeg fant heller ingen sammenheng mellom vekt og gjennomsnittshastighet jfr. 6.4.1 på side 92), men



Figur 4: Misledende korrelasjon 2

det er slik at mannlige langrennsløpere generelt presterer bedre og veier mer enn kvinnelige langrennsløpere. Vi ser av figur 4 at relasjonen til kjønn for henholdsvis vekt og gjennomsnittshastighet på hele løpet begge er positivt korrelerte jfr. plusstegnene, og uten tap av generalitet har jeg, som vist i figur 4, ikke lagt noen antagelse på årsakssammenhengen mellom kjønn og gjennomsnittshastighet på hele løpet, eller kjønn og vekt (jfr. linjene med pil på begge sider markert med + for positiv sammenheng). Den observerte positive sammenhengen mellom vekt og gjennomsnittshastighet på hele løpet kan altså forklares av kjønn.

5.1.1 Pearson's r og førsteordens partiell korrelasjon

En førsteordens partiellkorrelasjon gir sammenhengen mellom variablene X og Y etter at effekten fra den tredje variabelen Z er statistisk kontrollert for. Utledningen nedenfor forutsetter at X,Y,Z er multivariat normalfordelt. Under antagelsen om at Z er den eneste variabelen som påvirker variablene

Tabell 15: Fremgangsmåte for å finne 1.ordens partiellkorrelasjon mellom X og Y

X_i	Y_i	Z_i	$u_i = Y_i - 0.3 - 0.9Z_i$	$v_i = X_i - 1.2 - 0.6Z_i$
1	3	3	0	-2.0
2	1	2	-1.1	-0.4
3	2	1	0.8	1.2
4	4	4	0.1	0.4
5	5	5	0.2	0.8

X og Y, og at Z ikke blir påvirket av verken X eller Y, kan vi bruke den førsteordens partielle korrelasjonen, som den genuine korrelasjonen jfr. [34, p. 467-479], og korrelasjonen mellom X og Y uten å kontrollere for Z kalles illusorisk.
Jeg viser ved hjelp av et eksempel hvordan vi kan kontrollere for Z ved beregning av korrelasjonen mellom X og Y:
1) Hvert i'te individ sin verdi for Y(Y_i) kan beskrives ved en lineær funksjon av Z_i. Denne lineære funksjonen, uledet fra minste kvadraters metode, kan skrives som $Y_i = a_y + b_y Z_i + u_i$, der a_y og b_y er konstanter og u_i er residualet d.v.s. at u_i er den delen av Y_i, som ikke blir estimert av Z_i og det er ingen relasjon mellom Z og u slik at denne korrelasjonen vil være 0.
2)Tilsvarende, som under punkt 1), får vi at $X_i = a_x + b_x Z_i + v_i$, der a_x og b_x er konstanter og v_i er residualet og korrelasjonen mellom v og Z er altså 0.
3)Det er n par av (u_i, v_i), som representer de n observasjonene i utvalget. Hvert par av (u_i, v_i) er komponenten til X_i og Y_i, som ikke er estimert av Z_i. Derfor reflekterer korrelasjonen(r_{uv}) mellom u og v relasjonen mellom X og Y etter at effekten av Z er fjernet, eller vi kan si effekten av Z blir holdt konstant ved beregningen av korrelasjonen mellom X og Y. Dette er altså det vi definerer som en førsteordens partiell korrelasjon mellom X og Y når vi kontrollerer for Z.
Et numerisk eksempel er vist i tabell 15, der $u_i = Y_i - 0.3 - 0.9Z_i$ og $v_i = X_i - 1.2 - 0.6Z_i$ selvsagt bare er et resultat av de minste kvadraters løsninger på henholdsvis $Y_i = 0.3 + 0.9Z_i + u_i$ og $X_i = 1.2 + 0.6Z_i + v_i$. Vi kan så lett beregne r_{xy}=0.7, r_{xz}=0.6, r_{yz} = 0.9, r_{uz}=0, r_{vz}=0 og r_{uv}=0.46. Den førsteordens partielle korrelasjonen mellom X og Y statistisk kontrollert for Z kan også beregnes direkte ved formelen:

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{1 - r_{xz}^2}\sqrt{1 - r_{yz}^2}} \tag{135}$$

Med tallene fra eksempelet over ser vi at den førsteordens partielle korrelasjonskoeffisienten blir: $r_{xy.z} = \frac{0.7-0.6*0.9}{\sqrt{1-0.6^2}\sqrt{1-0.9^2}} = 0.46 = r_{u,v}$. For å teste nullhypotesen $\rho_{xy.z} = 0$ kan vi utføre t-testen med forventning 0 og standardfeil $\sqrt{\frac{1-r_{xy.z}^2}{n-3}}$ og antall frihetsgrader (n-3) jfr. [5]. Vi får dermed t = $\frac{r_{xy.z}\sqrt{n-3}}{1-r_{xy.z}^2} = \frac{0.46\sqrt{(5-3)}}{\sqrt{1-0.46^2}} = 0.73$, som er lavere enn den kritiske t-verdien på 4.30 for α=0.05 og tosidig test - vi får m.a.o. ikke forkastet nullhypotesen $\rho_{xy.z} = 0$.
For bevis se [3, 330-332] hvor man også enkelt generaliserer resultatet for førsteordens partiellkorrelasjon til p'te ordens partiell korrelasjon. Fort kan p'te ordens partiell korrelasjon oppsummeres som følger jfr. [35, s. 91-97] 1.utgave. Anta at H er settet av alle uavhengige variable $X_1, X_2, ..., X_n$, G_k er settet av alle de uavhengige variablene i settet H bortsett fra X_k, og Y er den avhengige variabelen. Vi finnner da den partielle korrelasjonen av X_k med Y ved å regressere X_k på G_k og Y på G_k for deretter å finne korrelasjonen mellom residualene, la oss si e_x og e_y, til de to regresjonene. Korrelasjonen mellom e_x og e_y kalles den partielle korrelasjonen(pr_k) mellom X_k og Y, fordi vi har fjernet effekten av de uavhengige variablene, som inngår i G_k fra både X_k og Y - fremgangsmåten er tilsvarende, som vi gjorde ovenfor når vi totalt bare hadde tre variable.

5.1.2 Kendall's t og førsteordens partiell korrelasjon

På liknende vis, som for Pearson's partielle korrelasjon diskutert ovenfor, kan vi oppdage Kendall's partielle rangkorrelasjon $t_{xy.z}$. Dersom vi har tre ordinale variabler X, Y og Z vil den partielle rang-korrelasjonen mellom X og Y hvor vi kontrollerer for Z kunne evalueres via:

$$t_{xy.z} = \frac{t_{xy} - t_{xz}t_{yz}}{\sqrt{1-t_{xz}^2}\sqrt{1-t_{yz}^2}} \tag{136}$$

Målefordelingen til $t_{xy.z}$ er ikke fordelt, som verken Kendall's t-fordelingen eller standardnormalfordelingen til z. Nullhypotesen($\tau_{xy.z} = 0$) kan testes via de kritiske verdier rapportert av Kendall og Gibbons [1, p. 237].

5.2 Semipartielle korrelasjoner

Resultatene i seksjonen her er hentet fra [35] 1.utgave.
I motsetning til partielle korrelasjoner, hvor ingen av variablene må oppfattes, som avhengig av de andre, vil semipartielle korrelasjoner typisk være aktuelt hvor vi har en avhengig variabel og flere uavhengige forklaringsvariable. I tilfelle med en uavhengig variabel X og avhengig variabel Y er den

semipartielle korrelasjonen identisk med korrelasjonen mellom de to variablene X og Y. En semipartiell korrelasjon(sr) indikerer det unike bidraget av en uavhengig variabel. Spesielt forteller sr^2 for en variabel hvor mye R^2(se seksjon 7 på side 112 for definisjon) vil øke eller synke med dersom den fjernes fra regresjonslikningen. Anta, jfr. de semipartielle korrelasjonsanalysene i seksjonene 6.4.4 på side 100 og 6.4.7 på side 106, at vi har uavhengige variable X_1(hastighet opp), X_2(hastighet flatt), X_3(hastighet ned) og avhengig variabel Y(hastighet totalt), og anta at variablene er standardiserte. For å få den semipartielle korrelasjonen av f.eks X_1 med Y utfører vi en regresjon av X_1 på X_2 og X_3. Residualet fra denne regresjonen(altså forskjellen mellom den predikerte verdien av X_1 og den faktiske verdien til X_1) er e_1. Den semipartielle korrelasjonen er da korrelasjonen mellom e_1 og Y. Det kalles en semipartiell korrelasjon, fordi effekten av X_2 og X_3 har blitt fjernet fra X_1, men ikke fra Y. Noen sammenhenger mellom semipartielle korrelasjoner og R^2 er:

- Jo mer tolerant en variabel er (jo mindre den er korrelert med de andre variablene) jo større vil det unike bidraget til R^2 være.
- Med en gang en variabel legges til, eller fjernes fra regresjonsligningen vil alle de andre semipartielle korrelasjonene kunne endres. De semipartielle korrelasjonene forteller bare om endringer i R^2 for en variabel av gangen!.
- Semipartielle korrelasjoner er typisk benyttet i stegvis regresjonsprosedyre, for å avgjøre hvilke variabler, som skal inngå i den endelige regresjonslikningen. I en foroverrettet stegvis regresjonsprosedyre vil den variabelen, som gir det største bidraget til R^2(altså den variabelen som har størst semipartiell korrelasjon), bli lagt til som neste variabel forutsatt at den er signifikant. I en bakoverrettet stegvis regresjonsprosedyre vil variabelen, som produserer den minste økningen i R^2(typisk den med lavest semipartiell korrelasjon), være neste som droppes forutsatt at den ikke er signifikant.

Videre har vi dersom Y er den avhengige variabelen, H er settet av alle uavhengige variable, og G_k er settet av alle uavhengige variable bortsett fra variabel X_k, følgende sammenheng:
$R^2_{YH} = R^2_{YG_k} + sr^2_k <=> R^2_{YG_k} = R^2_{YH} - sr^2_k$ jfr. [35, s. 91-97] 1.utgave. M.a.o. når Y er regressert på alle uavhengige variable X bortsett fra X_k er R^2 lik kvadratet av korrelasjonen til Y regressert på alle X bortsett fra X_k pluss kvadratet av den semipartielle korrelasjonen for X_k. Ønsker vi å vite hva R^2 vil være hvis f.eks X_k fjernes fra likningen trekker vi bare sr^2_k fra R^2_{YH}.

Forøvrig har vi følgende sammenhenger mellom partielle og semipartielle korrelasjoner: $sr_k^2 = \frac{pr_k^2}{1-pr_k^2}(1-R_{YH}^2)$ og $pr_k^2 = \frac{sr_k^2}{1-R_{Y_{G_k}}^2}$, der pr_k^2 referer til kvadratet av den partielle korrelasjonen definert ovenfor under 5.1.1 på side 81.
Jeg vil i analysene nedenfor typisk benytte semipartielle korrelasjoner når jeg diskuterer pacingstrategi hos langrennsløperne vi studerer, jfr. de semipartielle korrelasjonsanalysene i 6.4.4 på side 100 og 6.4.7 på side 106.

5.3 Pearson's r og aggregerte målinger

Størrelsen på Pearson's r kan også bli påvirket når den beregnes ved en kombinasjon av flere 'ulike' målinger. Tenk deg at en rekke kvinnelige utøvere i langrenn fra de tre landene Norge, Sverige og Danmark blir testet for overall fitness av hvert av sine forbund og ikke lenge etterpå, på de samme variablene, samlet blir testet av et internasjonalt forbund. Korrelasjonen mellom det internasjonale forbundet og de enkelte nasjonale forbund vil typisk alle ha relativt høy positiv korrelasjon i tillegg til at det er naturlig å tenke seg at de tre korrelasjonene er nokså like. Derimot hvis vi ser på alle utøverne samlet risikerer vi å få en negativ korrelasjon mellom overall fitness score beregnet av henholdsvis det internasjonale forbundet og de nasjonale forbund. Overall fitness score for utøverene fra det enkelte land kan skille på den måten at gruppen fra Norge generelt får lave score av det nasjonale forbundet smln. med høye score av det internasjonale forbundet, gruppen fra Sverige får typisk medium score både fra nasjonalt og internasjonalt forbund mens gruppen fra Danmark generelt får høye score fra nasjonalt forbund og lave score fra internasjonalt forbund. Samlet gir dette en negativ korrelasjon mellom overall fitness score beregnet av det internasjonale forbund og de nasjonale forbund. Årsaken til dette er at de nasjonale forbundene bruker ulik scala på overall fitness score. Dette er nok et eksempel på Simpsons paradoks, som vi studerte i 5.1 på side 79, og vi konstaterer at når vi beregner korrelasjonskoeffisienter basert på en kombinasjon av ulike målinger, vil denne korrelasjonen fort være misvisende i forhold til hva som er den faktiske relasjonen.

5.4 Kort om restriksjoner på hvilke verdier en variabel kan ta

Korrelasjonen mellom to variable kan også bli påvirket av at man begrenser skalaen for hvilke verdier en, eller begge variablene kan ta. Man kan begrense skalaen i den ene enden, eller i begge ender. Det er mange årsaker til at målinger på en variabel blir begrenset i hvilke verdier de kan ta på en skala:

- Måleinstrumentene er f.eks. ikke sensitive nok slik at ikke alle verdier kan måles.
- Mennesker vil ikke alltid besvare spørsmål korrekt, som f.eks. alkoholinntak og rusmisbruk.
- En egenskap i en populasjon kan være skjevt fordelt
- Dersom en studerer en relativt homogen grupppe vil studieobjektene ha enkelte likhetstrekk. Studieobjektene vil score nokså likt på enkelte variable, og vil altså være mye likere på disse bakgrunnsvariablene sammenlignet med en tilfeldig måling. For eksempel, hvis en tredje variabel, som høy VO2maks, er relatert til gode resultater i kondisjon og til gode prestasjoner i styrke, vil korrelasjonen mellom gode resultater i kondisjon og gode resultater i styrke, i en gruppe mennesker med høy VO2maks, tendere mot å være forskjellig sammenlignet med et utvalg fra en tilfeldig gruppe mennesker.

En relativt enkel formel for hvordan man kan korrigere for skalarestriksjoner ved bruk av Pearson's r finner du i [5, p. 58].

5.5 Kort om målefeil

Tilfeldige målefeil, som et ukalibrert måleinstrument, eller en sliten person, som tester, er en viktig faktor som kan påvirke korrelasjonen. Det er nødvendig å bruke gyldige og pålitelige måleinstrumenter. Hvis det er en stor andel av tilfeldige feil i målingene blir disse målingene mye mindre pålitelige/til å stole på. En relativt enkel formel, som kan korrigere for mangel på pålitelighet, ved bruk av Pearson's r, finner du i [5, p. 68].

6 Analyse av data på Norges beste langrennsutøvere

6.1 Data

Data, som blir benyttet i analysene nedenfor, ble blant annet samlet inn under en internasjonal konkurranse på Beitostølen 22-23 November 2013(heretter referert til som Beitodataene2013). Det var to individuelle konkurranser - en i klassisk og en i skøyting. Konkurransene foregikk i samme 5 km-sløyfe og slik at kvinnene gikk 10 km og herrene 15km begge dager. Løypa ble

kartlagt ved hjelp av GPS og et barometer for å få en gyldig løype og høydeprofil. Tid, hastighet og hjertefrekvens for oppover-, flatt- og nedoverterreng ble beregnet ved at utøverne gjennom hele løpet bar en GPS- og hjertefrekvensmonitor.
Beitodataene2013 omfatter 9 kvinner klassisk og 6 kvinner skøyting, der 5 gikk begge dager, og 7 menn klassisk og 8 menn skøyting, der 4 gikk begge dager.
Tilsvarende konkurranser og datainnsamling, som den i 2013, ble gjennomført 21-22 november 2014. Det var kun data på kvinnene(heretter kalt Beitodataene2014), som var gode nok til å utføre statistiske undersøkelser på denne gang. Jeg har data på 10 kvinner klassisk og 9 kvinner skøyting, der 7 gikk begge dager, men hvor vi har 'fulle' data kun på 6.
I tillegg fikk jeg endel labdata på kvinnene, som gikk i 2014. I analysene nedenfor har jeg av labdataene benyttet meg av variablene høyde i cm, vekt i kg, body mass index(bmi(vekt/høyde$(m)^2$)), VO2maxLØP(maksimalt oksygenopptak funnet ved løpetest på tredemølle målt i mL/(kg*min)) , VO2maxDIA(VO2max funnet ved 3min full innsats i diagonalgang), VO2maxSTA(VO2max funnet ved 3min full innsats i staking), laktatmaxDIA (laktat målt i mmol/l etter 3 min maks innsats i diagonalgang), laktatmaxSTA(laktat målt i mmol/l etter 3 min maks innsats i staking).
I seksjonen her vil jeg i liten grad benytte meg av de statistiske undersøkelsene jeg gjorde i forbindelse med prosjektet nevnt innledningsvis i seksjon 1.1 på side 6, men vil i oppgaven her kun fokusere på parametriske og ikke-parametriske korrelasjonsmål. Jeg vil diskutere valg av ulike typer korrelasjonsmål brukt på Beitodataene fra 2013 og 2014 i tillegg til de nevnte labdataene.

6.2 Aktuelle korrelasjonsmål og teknikker for å utføre inferens på data

I analysene nedenfor vil jeg benytte metodene 1-8 og 11-14 definert i appendiks 1 i A på side 117. De første 8 metodene definerer ulike måter å teste nullhypotesen om at det ikke er en sammenheng/korrelasjon mellom to variable, og der det er mulig, beregnes konfidensintervaller for den aktuelle populasjonskorrelasjonsparameteren. Metode 11 blir brukt til å beregne partielle korrelasjoner. Metode 12 og 13 avgjør hvorvidt korrelasjoner fra to uavhengige utvalg er signifikant forskjellige. Metode 14 beregner semipartielle korrelasjoner.
Kort om hvordan metodene 1-8 benyttes i analysene:

Metode 1(eksakt målefordeling til Pearson r) er eksakt når data kommer fra en bivariat normalfordeling. Jeg bemerker at selvom vi ikke kan forkaste hypotesen om at data er normalfordelt, er det likevel ingen garanti for at data er normalfordelt. Spesielt blir normalfordelingsantagelsen diskutabel når data er få. På den annen side hvis vi får en klar indikasjon på at data er normalfordelt, i tillegg til at det ikke er klare holdepunkter for at variablene vi studerer har egenskaper ved seg, som bør tyde på at populasjonsfordelingen ikke er en bivariatnormalfordeling, er metoden god som modell.
Metode 2(Fishertransformasjon av Pearson's r) vil ikke være aktuell å bruke på data her grunnet få observasjoner og blir ikke kommentert i analysene, men slik at p-verdier og konfidensintervaller blir oppgitt for den interesserte leser.
Metode 3(permutasjonsfordeling) er eksakt i bruk når vi ønsker å teste om vår observerte Pearson's r gir grunnlag for å hevde at populasjonsparameteren ρ er forskjellig fra 0. Metoden er helt uavhengig av fordeling på data.
Metode 4(tilnærmer permutasjonsfordelingen med t-fordelingen) er en god metode, uavhengig av fordeling når estimatene for g'ene er nær null(se 3.1.2 på side 42), og vi ønsker å teste om vår observerte Pearson's r er signifikant forskjellig fra null. Ved bruk av metoden vil det alltid være en usikkerhet knyttet til hvorvidt de empirisk beregnede g'ene gir gode estimater for de tilsvarende g'ene i populasjonen.
Metode 5(ikke-parametrisk bootstrapping) - vi antar at alle observerte par av to variable X,Y har lik sannsynlighet for å forekomme og tester på denne bakgrunn H_0: ρ=0. Jo lenger 0 er fra konfidensintervallet vi lager jo sterkere blir konklusjonen når vi tester H_0. Grunnet antagelsene vi har lagt på den kumulative fordelingsfunksjonen til F vil denne metoden, ved fastsettelse av øvre og nedre grense, for det beregnede konfidensintervallet, kunne gi laveste grense lavere enn -1 og øverste grense større enn 1. I nevnte tilfeller erstatter jeg nedre og øvre grense med henholdsvis -1 og 1. Bruken av metoden her gjør det tilsynelatende lettere å lokalisere populasjonsparameteren sammenlignet med bruk av M1 ved at konfidensintervallene blir smalere. Årsaken til de smalere konfidensintervallene er at vi hevder å kjenne den kumulative fordelingsfunksjonen(med minimale antagelser) uavhengig av om vi har lite eller mye data. M1 tar hensyn til at vi har få observasjoner(Merk: benytter f.eks t-fordelingen dersom vi antar at ρ=0) i motsetning til M5(eller lettere interpreterbart M6 nedenfor), som beregner konfidensintervaller ved å trekke fra den antatte tilpassede fordelingen. M1 ville gitt langt smalere konfidensintervaller ved mye data mens mye data ikke påvirker konfidensintervallene beregnet ved M5 på samme måte.
Metode 6(parametrisk bootstrapping) vil bare benyttes hvor vi antar at fordelingen til X,Y følger en bivariat normalfordeling. Konfidensintervallene be-

regnet her sammenlignet med konfidensintervallene i M1 vil typisk være smalere. Årsaken til de smalere konfidensintervallene er at vi hevder å kjenne den kumulative fordelingsfunksjonen uavhengig av om vi har lite eller mye data. M1 tar hensyn til at vi har få observasjoner(Merk: benytter f.eks t-fordelingen dersom vi antar at ρ=0) i motsetning til M6, som beregner konfidensintervaller ved å trekke fra den tilpassede normalfordelingen. M1 ville gitt langt smalere konfidensintervaller ved mye data mens mye data ikke påvirker konfidensintervallene beregnet ved M6 på samme måte. Under antagelsen om bivariat normalfordeling vil jeg av nevnte grunner, som den klare hovedregel, alltid la M1 være superior M6 og velger derfor i stor grad ikke å kommentere M6 i analysene nedenfor, men oppgir konfidensintervallene for den interesserte leser. Tilsvarende, som i metode 5, vil det ved fastsettelse av øvre og nedre grense for konfidensintervallet kunne forekomme at |øvre| og/eller |nedre| grense er større enn 1. I nevnte tilfeller lar jeg nedre grense aldri være lavere enn -1, eller øvre grense høyere enn 1.
Metodene 7 og 8 (Kendall's t og Spearman's r_s) er eksakte og robuste da de kan benyttes uavhengig av fordeling og ikke er sensitive for uteliggere. Grunnet min lille tilgang på data gir bruk av disse korrelasjonsmålene kun opphav til å forkaste eller godta H_0:$\tau = 0$, H_0:$\rho_s = 0$. Metodene 7 og 8 er ikke helt sammenlignbare med de foregående metodene da vi her ikke måler lineær sammenheng i en Pearson's r verden, men kun monotonisiteten til variablene.
Dersom vi ikke har en underliggende bivariat normalfordeling vil bare de ikke-parametriske metodene 3,4,5,7 og 8 være aktuelle.
Jeg vil i resten av seksjon 6 bruke M1,M2..,M8, som en måte å referere til metodene 1-8. Jeg gjør dette i stedenfor, det mer omstendelige, å referere til den eksakte fordelingen til Pearson's r gitt en underliggende bivariat normalfordeling(M1), normalfordelingen etter Fishertransformasjon gitt en underliggende bivariat normalfordeling(M2), permutasjonsfordelingen ved beregning av Pearson's r(M3), t-fordelingen(som tilnærming til permutasjonsfordelingen ved beregning av Pearson's r)(M4), ikke-parametrisk bootstrapping ved beregning av Pearson's r(M5), parametrisk bootstrapping med antatt underliggende bivariat normalfordeling(M6), permutasjonsfordelingen til Kendall's t(M7), eller til permutasjonsfordelingen til Spearman's r_s(M8).

6.3 Bruk av korrelasjonsmål i toppidretten

6.3.1 n liten

Generelt har vi sett at det er vanskelig å si noe statistisk signifikant for en populasjon når utvalget er lite og spesielt er dette tilfelle når data ikke stam-

mer fra en bivariat normalfordeling.
Ved statistiske undersøkelser vil det ofte være viktig at gruppen vi undersøker er homogen jfr. diskusjonen i 5.4 på side 85 for at våre resultater skal gi gyldighet nettopp til denne gruppe individer. Dette er forklaringen på at vi ikke alltid kan ta med de nest beste, eller mosjonistene når vi studerer egenskaper ved Norges beste utøvere og at data i toppidretten ofte vil være få. Et eksempel kan være analyse på verdens beste kvinnelige hoppere. For at gruppen skal være nogenlunde homogen definerer vi verdens beste kvinnelige hoppere til dem, som gjennom en sesong, ligger innenfor 90% av kapasiteten til verdenscuplederen. Tilgangen på individer vil kanskje være maksimalt n=5 og vi ville i en slik analyse typisk studert alle 5 hopperne og slik at vi per beregningstidspunkt altså ikke har tilgang på flere individer.
Enkelte ganger kan det være at vi kan få n høyere ved å samle data på tvers av nasjonene. Et problem kan da være at nasjonene ikke vil utveksle informasjonen de har om sine løpere og n kan ikke bli høyere enn den homogene gruppen den enkelte nasjon har tilgang på. Langrennsdataene jeg benytter kan illustrere problemstillingen. Hadde jeg hatt data på tilsvarende gode utenlandske løpere, som de norske, ville vi nok kunne opprettholdt homogeniteten også for n=20-30.
En måte å omgå problemet med lite data er at man samler data på utøvere over tid og på denne måten få n høyere - ofte vil man da oppdage at nytt og bedre utstyr, nye teknikker m.m. gjør mye av dataene ubrukelige.
Alle konklusjonene knyttet til de konkrete analysene nedenfor, og konklusjoner på tilsvarende analyser på lite data, vil og bør undersøkes nærmere ved hjelp av flere tilsvarende undersøkelser på lite data, eller av en undersøkelse på mye data, før de f.eks. kan benyttes i treningsfilosofier, ved seleksjon, eller ved anbefalinger for hvorvidt en utøver bør satse på en idrett, som levebrød eller ikke.

6.3.2 Linearitet eller monotonisitet. Årsakssammenheng.

Ved valg av korrelasjonsmål er spørsmålet hvilket som egner seg best. Er det en linearitet, eller en monotonisitet vi ønsker å måle, eller ingen av delene? I oppgaven her studerer vi kun eventuelle lineære og monotonistiske sammenhenger.
Er det uteliggere har vi sett at det ofte er fornuftig å velge monotonisitet og altså Spearman's r_s, eller Kendall's t. Dersom det ikke er klare holdepunkter for at det er naturlig at variablene skal ha en lineær sammenheng, men at monotonisiteten ønskes målt, velger vi også Spearman's r_s eller Kendall's t.
I toppidretten måler vi ofte variable opp mot sluttresultat i idrettsgrenen. Utøvernes fokus på sluttresultatet er svært ofte f.eks. ikke en bestemt tid, men

posisjon/rangering. Dersom en eller to av variablene har et rangeringsfokus vil det beste korrelasjonsmålet være Spearman's r_s eller Kendall's t med mindre man mener det likevel er linearitet man måler.
Ved testing av hypotesene $H_1, H_2, ..., H_{14}$, definert i seksjon 6.4 nedenfor, forutsetter jeg ved interpretasjonen av eventuelt beregnede statistiske signifikante korrelasjoner at det eksisterer årsakssammenheng mellom de aktuelle variablene uten at jeg selv har forsket på dette. Jeg bemerker, klart og tydelig, at sistnevnte antagelse er svært utsatt for feil jfr. diskusjonen i seksjon 5 på side 78 om misvisende og illusorisk korrelasjon.

6.4 Hypotesene

Jeg vil i seksjonen her teste følgende nullhypoteser:
H_1 = Det er ingen sammenheng mellom vekt og hastighet i løpet(på hele løpet og i de enkelte seksjoner opp, flatt og ned)for kvinner klassisk/skøyting i 2014.
H_2 = Det er ingen sammenheng mellom høyde og hastighet i løpet(på hele løpet og i de enkelte seksjoner opp, flatt og ned) for kvinner klassisk/skøyting i 2014.
H_3 = Det er ingen sammenheng mellom bmi og hastighet i løpet(på hele løpet og i de enkelte seksjoner opp, flatt og ned) for kvinner klassisk/skøyting i 2014.
H_4 = Det er ingen sammenheng mellom VO2maxDIA og hastighet i løpet(på hele løpet og i de enkelte seksjoner opp, flatt og ned) for kvinner klassisk/skøyting i 2014.
H_5 = Det er ingen sammenheng mellom VO2maxSTA og hastighet i løpet(på hele løpet og i de enkelte seksjoner opp, flatt og ned) for kvinner klassisk/skøyting i 2014.
H_6 = Det er ingen sammenheng mellom VO2maxSTA/VO2maxDIA og hastighet i løpet(på hele løpet og i de enkelte seksjoner opp, flatt og ned) for kvinner klassisk/skøyting i 2014.
H_7 = Det er ingen sammenheng mellom laktatmaxDIA og hastighet i løpet(på hele løpet og i de enkelte seksjoner opp, flatt og ned) for kvinner klassisk/skøyting i 2014.
H_8 = Det er ingen sammenheng mellom laktatmaxSTA og hastighet i løpet(på hele løpet og i de enkelte seksjoner opp, flatt og ned) for kvinner klassisk/skøyting i 2014.
H_9 = Det er ingen sammenheng mellom gjennomsnittshastighet på løpet og hastighet i de enkelte seksjoner(opp, flatt og ned) for kvinner klassisk/skøyting i 2014.
H_{10} = Det er ingen sammenheng mellom gjennomsnittshastighet på løpet og

gjennomsnittspuls totalt, eller gjennomsnittspuls i de enkelte seksjoner(opp, flatt og ned) for kvinner klassisk/skøyting i 2014.
H_{11} = Det er ingen sammenheng mellom gjennomsnittshastighet på løpet og gjennomsnittspuls i de enkelte seksjoner(opp, flatt og ned)/gjennomsnittspuls totalt for kvinner klassisk/skøyting i 2014.
H_{12} = Det er ingen sammenheng mellom gjennomsnittshastighet på løpet og hastighet i de enkelte seksjoner(opp, flatt og ned)for kvinner klassisk/skøyting i 2013.
H_{13} = Det er ingen sammenheng mellom gjennomsnittshastighet på løpet og gjennomsnittspuls totalt, eller gjennomsnittspuls i de enkelte seksjoner(opp, flatt og ned) for kvinner klassisk/skøyting i 2013.
H_{14} = Det er ingen sammenheng mellom gjennomsnittshastighet på løpet og gjennomsnittspuls i de enkelte seksjoner(opp, flatt og ned)/gjennomsnittspuls totalt for kvinner klassisk/skøyting i 2013.
H_{15} = Korrelasjonen mellom VO2maxDIA og gjennomsnittshastighet på løpet kvinner 2014 skøyting er misvisende når vi kontroller for antall treningstimer.
H_{16} = Korrelasjonen mellom VO2maxSTA og gjennomsnittshastighet på hele løpet kvinner 2014 skøyting er misvisende når vi kontroller for antall treningstimer.
I seksjonene 6.4.1 til 6.4.6 på side 105 ser vi på Beitodataene2014.
I seksjonene 6.4.7 på side 106 til 6.4.9 på side 110 ser vi på Beitodataene2013.
I seksjonen 6.4.10 på side 110 kommenteres raskt at sammenligning av korrelasjoner beregnet i to uavhengige utvalg, relatert til mine data, ikke kan utføres statistisk forsvarlig.
I seksjonen 6.4.11 på side 110 tester vi H_{15} og H_{16} hvor vi igjen benytter Beitodataene2014.

6.4.1 Hastighet sammenlignet med vekt(H_1), høyde(H_2) og bmi(H_3) - 2014 data.

Jeg fant ingen sammenheng verken for vekt eller bmi med gjennnomsnittshastighet på 10km klassisk eller 10km skøyting i de ulike segmentene opp, bort og ned, eller på hele distansen. Det er kun for variabelen høyde jeg kan forkaste hypotesen(H_2) om at det ikke er noen sammenheng med gjennomsnittshastighet.

Høyde sammenlignet med hastighet på hele løpet og de enkelte segmenter opp, flatt og ned(H_2) jfr. tabell 16 på side 94.

Normalfordelte data:
Kontroll av den bivariate normalantagelsen(shapiro wilk test) for høyde og gjennomsnittshastighet for henholdsvis hele løypa, oppoversegmentet, bortoversegmentet og nedoversegmentet ga p-verdiene 0.44,0.36,0.21,0.66 og 0.29,0.18,0.35,0.35 for respektive klassisk og skøyting. Det er dermed ikke for noen segmenter grunnlag for å forkaste hypotesen om at den bivariate normalantagelsen holder. Dette er likevel på langt nær noen garanti for at den bivariate normalantagelsen holder og dermed at de aktuelle variablene har en lineær sammenheng. Det er forøvrig ikke klare tegn til uteliggere og siden den bivariate normalantagelsen holder vil det ikke være feil å benytte den parametriske metoden M1. De tildels svake p-verdiene innebærer likevel at det kan være fornuftig å legge mer vekt på de ikke-parametriske metodene 3,4,5,7,8 nedenfor, sammenlignet med den aktuelle parametriske metodene 1(og 6).

Parametriske korrelasjonsmål:
M1 gir signifikant resultat totalt i klassisk og svært nær signifikant resultat totalt i friteknikk. Jeg fikk signifikante resultater oppover i begge stilarter. Det er uansett, som vi ser av konfidensintervallene, svært vanskelig å lokalisere populasjonsparameteren ρ. I flatseksjonen og nedoverseksjonen er det også en negativ korrelasjon, men ingen signifikante resultater.

Ikke-parametriske korrelasjonsmål:
M3 gir signifikante resultater totalt og oppover i klassisk og svært nær signifikante resultater totalt og oppover i skøyting. I flatseksjonen og nedoverseksjonen er det også en negativ korrelasjon, men ingen signifikante resultater.
M3 gir tilsvarende resultater, som M1, men gir ikke mulighet for å studere konfidensintervaller.
Begrunnelsen for å kunne benytte M4 til å teste hvorvidt H_0:ρ=0 var at E(r^3) og E(r^4) begge var nær 0. Vi ser at E(r^3) er nær 0, men at E(r^4) er en anelse høyere enn null, men konstaterer at M4 gir de samme konklusjoner som M3.
M5 gir en tydelig indikasjon på at det foreligger en signifikant sammenheng mellom høyde og gjennomsnittshastighet på hele løpet i begge teknikker. Det må likevel, som påpekt i seksjon 6.2 på side 87, vises varsomhet med bruk av metoden når antall observasjoner er så få som her.
M7 og M8 gir generelt noe svakere korrelasjon(monotonistisk sammenheng mellom høyde og hastighet) enn Pearson's r og slik at det kun er i klassisk etter M7 opp, M8 totalt og opp, og i skøyting etter M8 opp, at vi har signifikante, eller nær signifikante resultater.

Konklusjon:
Det ser ut til å kunne være en sammenheng mellom høyde og hastighet og at denne gjør seg tydeligst gjeldende i oppoverseksjonen i beggge stilarter.

Tabell 16: Korrelasjon mellom høyde og gjennomsnittshastighet totalt og i de enkelte segmentene opp, flatt og ned - 2014

Metode M n=6	Observert korrelasjon klassisk/skøyting	P-verdi klassisk/skøyting	Konfidensintervall(0.95) klassisk/skøyting
M1 totalt	-0.828/-0.752	0.021/0.042	(-0.98,-0.22)/(-0.975,0.006)
opp	-0.89/-0.77	0.01/0.04	(-0.99,-0.46)/(-0.98,-0.04)
flatt	-0.6/-0.38	0.1/0.23	(-0.96,0.32)/(-0.92,0.59)
ned	-0.34/-0.51	0.25/0.15	(-0.91,0.62)/(-0.94,0.45)
M2 totalt	**	0.017/0.039	(-0.98,-0.09)/(-0.97,0.11)
opp	**	0.01/0.03	(-0.99,-0.33)/(-0.97,0.07)
flatt	**	0.11/0.24	(-0.95,0.38)/(-0.9,0.6)
ned	**	0.26/0.16	(-0.89,0.63)/(-0.93,0.49)
M3 totalt	**	0.019/0.04	
opp	**	0.01/0.03	
flatt	**	0.12/0.25	
ned	**	0.25/0.14	
M4' totalt	**	0.021/0.042	(-1,-0.05)/(-1,0.16)
opp	**	0.01/0.04	(-1,-0.26)/(-1,0.12)
flatt	**	0.1/0.23	(-1,0.51)/(-1,0.91)
ned	**	0.25/0.15	(-1,0.96)/(-1,0.69)
M7 totalt	-0.6/-0.47	0.07/0.14	
opp	-0.73/-0.6	0.03/0.07	
flatt***	-0.14(-0.13)/-0.33	0.36/0.23	
ned	-0.2/-0.33	0.36/0.23	
M8 totalt	-0.77/-0.71	0.05/0.07	
opp	-0.83/-0.77	0.03/0.05	
flatt****	-0.23/-0.6	0.33/0.12	
ned	-0.26/-0.37	0.33/0.25	
M5"	Korrelasjon	Konf.int. norm.app.	Konf.int. bootstrap
totalt	-0.828/-0.752	(-1,-0.39)/(-1,-0.23)	(-1,-0.66)/(-1,-0.51)
M6"'	Korrelasjon	Konf.int. norm.app.	Konf.int. bootstrap
totalt	-0.828/-0.752	(-1,-0.43)/(-1,-0.25)	(-1,-0.67)/(-1,-0.53)
M4'	Skjevhet: g1x, g1y	Kurtosis: g2x, g2y	Momenter: E(r^3), E(r^4)
totalt	-0.6,0.06/-0.6,0.04	-0.44,-1.43/-0.44,-1.64	-0.001,0.09/-0.001,0.09
opp	-0.6,0.15/-0.6,0.07	-0.44,-1.64/-0.44,-1.82	-0.002,0.09/-0.001,0.09
flatt	-0.6,0.006/-0.6,0.11	-0.44,-1.28/-0.44,-0.76	-0.0001,0.09/-0.002,0.09
ned	-0.6,-0.4/-0.6,-0.29	-0.44,-0.75/-0.44,-1.01	0.007,0.09/0.005,0.09
M5"	Skjevhet $\hat{\rho}$	Standardavvik $\hat{\rho}$	
totalt	0.01/-0.01	0.22/0.27	
M6"'	Skjevhet $\hat{\rho}$	Standardavvik $\hat{\rho}$	
totalt	0.03/0.04	0.2/0.26	

**Identisk med M1.

***$t_b(t_a)$ ved ties og p-verdi beregnes da ved normaltilnærming med kontinuitetskorreksjon.

****r_{sb} ved ties og p-verdi beregnes da ved normaltilnærming med kontinuitetskorreksjon.

I nedover- og flatterrenget eksisterer ikke signifikante sammenhenger. Sammenhengen kan både oppfattes som lineær, og monotonistisk, men slik at M7 bare såvidt gir signifikans oppover i klassisk og ellers ikke.
Sammenligner vi stilartene er den negative tendensen klarest i klassisk totalt, oppover og bortover. Nedover er den negative tendensen klarest i skøyting - årsaken til dette kan være at man som høy, relativt sett, fanger mer vind i skøyting enn i klassisk, når man sammenligner nedover med oppover- og bortoverterreng.
Vi kan konkludere med at det ser ut til å være negativt å være høy for å gå fort på ski i oppoverterreng, spesielt i klassisk, dersom du er kvinne.

6.4.2 Hastighet sammenlignet med laktatmaxDIA(H_7) og laktatmaxSTA(H_8) - 2014-data

Laktat, eller melkesyre, er det normale endeproduktet ved anaerob nedbrytning av glukose. Med tilstrekkelig oksygen tilstede vil glukose nedbrytes til karbondioksyd og vann. Laktat dannes når det ikke er tilstrekkelig oksygen tilstede i denne nedbrytningen. Med oksygen tilstede vil laktat oksyderes til pyruvat(pyrodruesyre). Pyruvat kan forbrennes videre ved flere mekanismer. Ved å måle laktatkonsentrasjonen i blodet ved økende belastning vil vi kunne finne ut hvor hardt du kan arbeide før produksjonen av laktat overstiger eliminasjonen. Denne belastningen er best kjent som anaerob terskel og sier noe om hvor stor andel av ditt maksimale oksygenopptak du kan utnytte over tid. Oksygenopptaket og ikke minst evnen til å nyttigjøre seg oksygenet man tar opp vil selvsagt, med bakgrunn i definisjonen ovenfor, være to faktorer, som vil være med på å avgjøre laktatnivået.

Konklusjon:
Jeg fant ingen signifikante sammenhenger verken mellom mengde laktat oppnådd etter tre minutter maks innsats på diagonalgang, eller staking og gjennomsnittshastighet. Det var dog en svak tendens mot lavere gjennomsnittshastighet ved høye laktatverdier i staking, men ingen signifikante sammenhenger.

6.4.3 Hastighet sammenlignet med VO2maxDIA(H_4), VO2maxSTA(H_5) og VO2maxSTA/VO2maxDIA(H_6) - 2014-data

VO2max diagonalgang(H_4) jfr. tabell 17 på neste side

Normalfordelte data:
Kontroll av den bivariate normalantagelsen(shapiro wilk test) forkastet hy-

Tabell 17: Korrelasjon mellom VO2max diagonalgang og gjennomsnittshastighet totalt og i de enkelte segmenter opp, flatt og ned - 2014

Metode M n=6	Observert korrelasjon klassisk/skøyting	P-verdi klassisk/skøyting	Konfidensintervall(0.95) klassisk/skøyting
M1 totalt	0.96/0.88	0.001/0.01	(0.78,0.997)/(0.43,0.99)
opp	0.95/0.85	0.001/0.02	(0.75,0.996)/(0.31,0.99)
flatt	0.83/0.62	0.02/0.1	(0.24,0.98)/(-0.29,0.96)
ned	0.65/0.78	0.08/0.03	(-0.24,0.96)/(0.07,0.98)
M2 totalt	**	0.0002/0.01	(0.7,0.995)/(0.3,0.99)
opp	**	0.0003/0.01	(0.66,0.99)/(0.18,0.99)
flatt	**	0.02/0.1	(0.11,0.98)/(-0.35,0.95)
ned	**	0.08/0.03	(-0.31,0.95)/(-0.05,0.97)
M3 totalt	**	0.003/0.01	
opp	**	0.001/0.02	
flatt	**	0.03/0.09	
ned	**	0.07/0.03	
M4' totalt	**	0.001/0.01	(0.57,1)/(0.24,1)
opp	**	0.002/0.02	(0.54,1)/(0.13,1)
flatt	**	0.02/0.1	(0.07,1)/(-0.47,1)
ned	**	0.08/0.03	(-0.42,1)/(-0.09,1)
M7 totalt	0.87/0.73	0.008/0.03	
opp	1/0.87	0/0.01	
flatt***	0.41(0.4)/0.6	0.14/0.07	
ned	0.47/0.6	0.14/0.07	
M8 totalt	0.94/0.89	0.01/0.02	
opp	1/0.94	0/0.01	
flatt****	0.55/0.77	0.15/0.05	
ned	0.54/0.66	0.15/0.09	
M5"	Korrelasjon	Konf.int. norm.app.	Konf.int. bootstrap
totalt	0.96/0.88	(0.91,1)/(0.56,1)	(0.92,1)/(0.77,1)
M6"'	Korrelasjon	Konf.int. norm.app.	Konf.int. bootstrap
totalt	0.96/ 0.88	(0.83,1)/(0.58,1)	(0.92,1)/(0.78,1)
M4'	Skjevhet: g1x, g1y	Kurtosis: g2x, g2y	Momenter: E(r^3), E(r^4)
totalt	0.49,0.06/0.49,0.04	-0.74,-1.43/-0.74,-1.64	0.001,0.09/0.001,0.09
opp	0.49,0.15/0.49,0.07	-0.74,-1.64/-0.74,-1.82	0.002,0.09/0.001,0.09
flatt	0.49,0.006/0.49,0.11	-0.74,-0.13/-0.74,-0.76	0.0001,0.09/0.001,0.09
ned	0.49,-0.4/0.49,-0.29	-0.74,-0.75/-0.74,-1.01	-0.01,0.09/-0.004,0.09
M5"	Skjevhet $\hat{\rho}$	Standardavvik $\hat{\rho}$	
totalt	0.02/0.02	0.03/0.17	
M6"'	Skjevhet $\hat{\rho}$	Standardavvik $\hat{\rho}$	
totalt	-0.01/-0.03	0.07/0.15	

**Identisk med M1.

***$t_b(t_a)$ ved ties og p-verdi beregnes da ved normaltilnærming med kontinuitetskorreksjon.

****r_{sb} ved ties og p-verdi beregnes da ved normaltilnærming med kontinuitetskorreksjon.

Tabell 18: Korrelasjon mellom VO2max staking og gjennnomsnittshastighet totalt og i de enkelte segmenter opp, flatt og ned - 2014

Metode M n=6	Observert korrelasjon klassisk/skøyting	P-verdi klassisk/skøyting	Konfidensintervall(0.95) klassisk/skøyting
M1 totalt	0.84/0.71	0.02/0.06	(0.25,0.98)/(-0.11,0.97)
opp	0.81/0.65	0.03/0.08	(0.16,0.98)/(-0.23,0.96)
flatt	0.82/0.47	0.02/0.17	(0.2,0.98)/(-0.49,0.94)
ned	0.61/0.78	0.1/0.03	(-0.31,0.96)/(0.07,0.98)
M2 totalt	**	0.01/0.06	(0.12,0.98)/(-0.2,0.96)
opp	**	0.02/0.08	(0.04,0.98)/(-0.3,0.95)
flatt	**	0.02/0.18	(0.08,0.98)/(-0.52,0.92)
ned	**	0.1/0.03	(-0.36,0.95)/(-0.05,0.97)
M3 totalt	**	0.02/0.05	
opp	**	0.03/0.08	
flatt	**	0.03/0.18	
ned	**	0.08/0.025	
M4' totalt	**	0.02/0.06	(0.08,1)/(-0.27,1)
opp	**	0.03/0.08	(-0.0004,1)/(-0.4,1)
flatt	**	0.02/0.17	(0.03,1)/(-0.75,1)
ned	**	0.1/0.03	(-0.49,1)/(-0.09,1)
M7 totalt	0.6/0.47	0.07/0.14	
opp	0.73/0.6	0.03/0.07	
flatt***	0.41(0.4)/0.33	0.14/0.23	
ned	0.47/0.6	0.14/0.07	
M8 totalt	0.77/0.71	0.05/0.07	
opp	0.83/0.77	0.03/0.05	
flatt****	0.52/0.49	0.15/0.18	
ned	0.6/0.71	0.12/0.07	
M5"	Korrelasjon	Konf.int. norm.app.	Konf.int. bootstrap
totalt	0.84/0.71	(0.32,1)/(0.1,1)	(0.68,1)/(0.42,1)
M6"'	Korrelasjon	Konf.int. norm.app.	Konf.int. bootstrap
totalt	0.84/0.71	(0.46,1)/(0.15,1)	(0.69,1)/(0.45,1)
M4'	Skjevhet: g1x, g1y	Kurtosis: g2x, g2y	Momenter: E(r^3), E(r^4)
totalt	0.54,0.06/0.54,0.04	-0.66,-1.43/-0.66,-1.64	0.001,0.09/0.001,0.09
opp	0.54,0.15/0.54,0.07	-0.66,-1.64/-0.66,-1.82	0.002,0.09/0.001,0.09
flatt	0.54,0.006/0.54,0.11	-0.66,-0.13/-0.66,-0.76	0.0001,0.09/0.002,0.09
ned	0.54,-0.4/0.54,-0.29	-0.66,-0.75/-0.66,-1.01	-0.01,0.09/-0.004,0.09
M5"	Skjevhet $\hat{\rho}$	Standardavvik $\hat{\rho}$	
totalt	-0.05/-0.05	0.26/0.31	
M6"'	Skjevhet $\hat{\rho}$	Standardavvik $\hat{\rho}$	
totalt	-0.03/-0.04	0.2/0.28	

**Identisk med M1.

***$t_b(t_a)$ ved ties og p-verdi beregnes da ved normaltilnærming med kontinuitetskorreksjon.

****r_{sb} ved ties og p-verdi beregnes da ved normaltilnærming med kontinuitetskorreksjon.

potesen om en bivariat normalfordeling for VO2max diagonalgang og gjennomsnittshastighet på løpet sett under et og for alle seksjoner(opp, ned og flatt) både i klassisk og skøyting.
Jeg vil derfor, nedenfor, bare studere de ikke-parametriske metodene 3,4,5,7 og 8.
Korrelasjonsanalyse:
Metodene 3,4,7 og 8 ga alle signifikante resultater når man så distansen under et både i klassisk og skøyting. Når vi skiller på segmentene ser vi at signifikansen opprettholdes i oppoverterrenget. I flatseksjonen er det bare i klassisk ved bruk av metode 3 og 4 at vi har signifikans, eller nær signifikans(ellers en klar positiv tendens) - den lineære sammenhengen via Pearson r er altså tydeligere enn den rene monotonisiteten beregnet i metodene 7 og 8 i klassisk. I nedoverseksjonen er det bare i skøyting ved bruk av metode 3 og 4 at vi har en signifikant, eller nær signifikant sammenheng og igjen er monotonisiteten svakere enn den lineære sammenhengen målt via Pearson's r.
M4 ga tilsvarende verdier av E(r^3)og E(r^4), som i seksjon 6.4.1 på side 92 og gir også her tilsvarende resultater, som M3.
M5 gir en tydelig indikasjon på at det foreligger en signifikant sammenheng mellom VO2max diagonalgang og gjennomsnittshastighet på hele løpet i begge teknikker. Det må likevel, som påpekt i seksjon 6.2 på side 87, vises varsomhet med bruk av metoden når antall observasjoner er så få som her.
Konklusjon: Høy VO2max diagonalgang er viktig for å gå fort på ski både i klassisk og skøyting og spesielt viktig i oppoverseksjonen.
Sammenligner vi stilartene er det en tendens til at høy VO2max diagonalgang er viktigst i klassisk når man ser hele løpet under et og i oppoverseksjonen. Det samme gjelder i flatseksjonen ved bruk av metodene 3 og 4, men M7 og M8 viser en tydeligere tendens for skøyting i flatseksjonen. I nedoverseksjonen er det skøytedisiplinen, som gir høyest positiv korrelasjon etter de fire mest aktuelle metodene. Sistnevnte er et argument for at andre faktorer enn VO2max, som glid,teknikk m.m., i større grad er utslagsgivende for hastigheten i nedoverseksjonen.

VO2max staking(H_5) jfr. tabell 18 på forrige side

Normalfordelte data:
Test av den bivariate normalantagelsen forkastet ikke hypotesen om bivariat normalfordeling for VO2max staking og gjennomsnittshastighet løpet sett under et, eller for de enkelte seksjoner verken i klassisk eller skøyting. I klassisk og skøyting er p-verdiene for hele løpet, opp, ned og flatt respektive 0.4,0.65,0.22,0.73 og 0.28,0.42,0.12,0.13 etter en shapiro wilk test.

At vi ikke forkaster den bivariate normalantagelsen og dermed beholder muligheten av f.eks. å bruke M1(og M6) er likevel ingen garanti for at normalantagelsen holder og det vil være hensiktsmessig, med et så lite datamateriale, å også benytte ikke-parametriske tester ved avgjørelsen av om vi har signifikante sammenhenger, eller ikke.
De aktuelle metodene, for å studere hvorvidt det foreligger en signifikant korrelasjon mellom VO2max staking og gjennomsnittshastighet, er metodene 1,3,4,5,6,7,8.
Parametriske korrelasjonsmål:
M1 gir signifikant resultat totalt i klassisk og nært signifikant resultat totalt i friteknikk. Tilsvarende resultater oppover. I flatseksjonen har vi signifkant resultat i klassisk og en positiv tendens i skøyting. I nedoverseksjonen er det også en positiv tendens i klassisk, men signifikant sammenheng mellom hastighet og VO2max staking i skøyting. Det er uansett, som vi ser av konfidensintervallene, svært vanskelig å lokalisere populasjonsparameteren ρ.
Ikke-parametriske korrelasjonsmål:
M3 gir signifikant, eller nær signifikant resultat totalt, og signifikante, eller nær signifikante resultater oppover og flatt i klassisk. Det er klare positive tendenser totalt og oppover i skøyting mens det bortover bare er en positiv tendens. I nedoverseksjonen er det en klar positiv tendens i klassisk og signifikant resultat i friteknikk.
M4 ga tilsvarende verdier av E(r^4)og E(r^4), som i seksjon 6.4.1 på side 92 og gir også her tilsvarende resultater, som M3.
M5 gir en tydelig indikasjon på at det foreligger en signifikant sammenheng mellom VO2max staking og gjennomsnittshastighet på hele løpet i begge teknikker. Det må likevel, som påpekt i seksjon 6.2 på side 87, vises varsomhet med bruk av metoden når antall observasjoner er så få som her.
M7 og M8 gir generelt noe svakere korrelasjon(monotonistisk sammenheng) mellom VO2max staking og hastighet) enn Pearson's r og slik at M7 gir svakere tendens enn M8. Det er kun i klassisk etter M7 opp, M8 totalt og opp, og i skøyting etter M8 opp, at vi har signifikante, eller nær signifikante resultater. Jeg finner ikke umiddelbart noen grunn til at M7 skal være mer riktig å bruke enn M8, men for å minske sjansen for feilaktig å konkludere med signifikante resultater finner jeg det fornuftig å la M7 være superior M8.

Konklusjon:
Det ser ut til at høy VO2maxSTA er viktig, om i noe mindre grad enn VO2max diagonalgang, for å gå fort på ski i klassisk og skøyting, men slik at det var bare M1 ned, M3 ned og M8 opp, som ga signifikante resultater i skøyting. Det ser ut til at høy VO2maxSTA er viktigst i klassisk bortsett fra i nedoverseksjonen hvor typisk andre faktorer, som glid,teknikk m.m. i større

grad påvirker hastigheten enn i de øvrige seksjonene.
Vi konstaterer videre at både VO2maxSTA og VO2maxDIA begge er mer korrelert med hastigheter i klassisk enn i skøyting. Det kan dermed se ut til at VO2max-kapasiteten generelt er viktigere i klassisk enn i skøyting og at andre prestasjonsfremmede faktorer, som teknikk og arbeidsøkonomi kanskje er av større viktighet i skøytedisiplinen sammenlignet med klassiskdisiplinen. Dette kunne vært interessant å gjøre ytterligere analyser på.

VO2maxSTA i prosent av VO2maxDIA mot gjennomsnittshastighet totalt(H_6)

Konklusjon: Jeg fant ingen signifikante sammenhenger, og det var bare en anelse positiv korrelasjon både i skøyting og klassisk.
Det ser dermed ut som at det ikke er grunnlag for å hevde at høy VO2max i staking påvirker sluttresultatet på en annen måte enn høy VO2max i diagonalgang.

6.4.4 Gjennomsnittshastighet på hele løpet sammenlignet med hastighet i de ulike segmenter(H_9) - 2014-data jfr. tabell 19 på neste side

Normalfordelte data:
Test av den bivariate normalantagelsen forkastet ikke hypotesen om bivariat normalfordeling for gjennomsnittshastighet på hele løpet og gjennomsnittshastigheten i hver av segmentene(opp, flatt og ned) verken i klassisk eller skøyting. Enten vi ser på gjennomsnittshastighet for hele løpet, gjennomsnittshastigheten for den enkelte runde hver for seg, eller rundene samlet kan vi ikke forkaste hypotesen om en bivariat normalfordeling etter en shapiro wilk test.
Aktuelle metoder er derfor 1,3,4,5,6,7,8.
Korrelasjonsanalyse:
Det er en klart signifikant sammenheng mellom gjennomsnittshastighet oppover og gjennomsnittshastighet på hele løpet etter alle metoder og begge teknikker. I flatseksjonen har vi en signifikant sammenheng i skøyting etter alle metoder, mens i klassisk har vi signifikans, eller nær signifikans etter alle metoder bortsett fra M7 og M8. I nedoverterrenget er det i klassisk signifikans eller nær signifikans etter alle metoder bortsett fra M7 og M8 mens det i skøyting kun er etter M5 og M6 at vi har signifikante resultater. Det er ellers en klar tendens til positiv sammenheng i alle partier av løypa opp mot gjennomsnittshastighet på hele løpet.
Sammenligner vi de to skiteknikkene er det lite forskjell på sammenhengen

Tabell 19: Korrelasjon mellom gjennomsnitthastighet totalt og gjennomsnittshastighet i de ulike segmenter opp, flatt og ned - 2014

Metode M n=6	Observert korrelasjon klassisk/skøyting	P-verdi klassisk/skøyting	Konfidensintervall(0.95) klassisk/skøyting
M1			
opp	0.98/0.99	0.0002/0.0001	(0.89,0.998)/(0.92,0.998)
flatt	0.76/0.87	0.04/0.01	(0.01,0.98)/(0.35,0.99)
ned	0.76/0.7	0.04/0.06	(0.02,0.98)/(-0.13,0.97)
M2			
opp	**	0.00002/0.000004	(0.84,0.998)/(0.89,0.998)
flatt	**	0.04/0.01	(-0.1,0.97)/(0.22,0.98)
ned	**	0.04/0.06	(-0.09,0.97)/(-0.22,0.96)
M3			
opp	**	0.003/0.003	
flatt	**	0.03/0.003	
ned	**	0.04/0.06	
M4'			
opp	**	0.0002/0.0001	(0.71,1)/(0.76,1)
flatt	**	0.04/0.01	(-0.15,1)/(0.17,1)
ned	**	0.04/0.06	(-0.14,1)/(-0.29,1)
M7			
opp	0.87/0.87	0.008/0.008	
flatt***	0.55(0.53)/0.87	0.07/0.008	
ned	0.6/0.6	0.07/0.07	
M8			
opp	0.94/0.94	0.008/0.008	
flatt****	0.7/0.94	0.07/0.001	
ned	0.71/0.71	0.07/0.07	
M5"	Korrelasjon	Konf.int. norm.app.	Konf.int. bootstrap
opp	0.98/0.99	(0.79,1)/(0.76,1)	(0.96,1)/(0.97,1)
flatt	0.76/0.87	(0.06,1)/(0.6,1.13)	(0.52,1)/(0.73,1)
ned	0.76/0.7	(0.24,1)/(0.12,1)	(0.52,1)/(0.42,1)
M6"'	Korrelasjon	Konf.int. norm.app.	Konf.int. bootstrap
opp	0.98/0.99	(0.91,1)/(0.93,1)	(0.96,1)/(0.97,1)
flatt	0.76/0.87	(0.27,1)/(0.53,1)	(0.54,1)/(0.74,1)
ned	0.76/0.7	(0.27,1)/(0.14,1)	(0.55,1)/(0.43,1)
M4'	Skjevhet: g1x, g1y	Kurtosis: g2x, g2y	Momenter: E(r^3), E(r^4)
opp	0.15,0.06/0.07,0.04	-1.64,-1.43/-1.82,-1.64	0.0002,0.09/0.0001,0.09
flatt	0.006,0.06/0.11,0.04	-0.13,-1.43/-0.76,-1.64	0.00001,0.09/0.0001,0.09
ned	-0.4,0.06/-0.3,0.04	-0.75,-1.43/-1.01,-1.64	-0.001,0.09/-0.0003,0.09

**Identisk med M1.

***$t_b(t_a)$ ved ties og p-verdi beregnes da ved normaltilnærming med kontinuitetskorreksjon.

****r_{sb} ved ties og p-verdi beregnes da ved normaltilnærming med kontinuitetskorreksjon.

i oppover og nedoverseksjonen i de to disiplinene, men i flatseksjonen er sammehengen med gjennomsnittshastigheten på hele løpet klarest i skøytedisiplinen og korrelajonen i flatterrenget i skøytedisiplinen ligner mer på korrelasjonen i oppoverterrenget enn nedoverterrenget. Dersom man gjør disse funnene i ytterligere analyser vil det være et argument for at egenskapene man som skiløper benytter i oppoverseksjonen og flatseksjonen i skøyting er likere enn i klassisk.
Når jeg skiller på rundene finner jeg lite forskjeller sammenlignet med analysene ovenfor.

Korrelasjoner mellom de ulike segmentene - pacing:
Da alle seksjoner viser en klar positiv tendens med sluttresultatet er det lite som tyder på at løperne pacer ulikt i de ulike seksjoner. Det er likevel noe mindre positiv korrelasjon i nedoverterrenget og det tyder på at løpere, som f.eks. går fortest nedover ikke nødvendigvis er de samme som går fortest oppover. Når jeg studerer korrelasjonene mellom de ulike terrengtypene er disse klart positive og det er derfor ikke overraskende lite forskjeller mellom rangeringene av løperne relatert til hastighet i de ulike seksjoner. Rangeringene med 1 som dårligst og 6 som best på løperne (A,B,C,D,E,F) var i henholdsvis klassisk/skøyting totalt(1,2,3,5,4,6/1,2,3,5,6,4), opp(1,3,2,5,4,6/1,3,2,5,6,4), ned(1,2,5,6,3,4/1,2,5,4,3,6) og bort(1,2,4,5,3,6/1,2,3,6,5,4). Likevel ser vi at vinneren i skøyting bare har fjerde beste tid i nedoverseksjonen og nest beste i flatseksjonen og at vinneren i klassisk bare har tredje beste tid i nedoverseksjon, men ellers best.

Semipartielle korrelasjoner - pacing:
Ved å bruke lineær multippel regresjon herunder studere de semipartielle korrelasjonene jfr. 5.2 på side 83, konstaterer vi at oppoverseksjonen er den langt viktigste variabelen for å forklare og predikere gjennomsnittelig hastighet på hele løpet. Den semipartielle korrelasjon for oppover er 0.561/0.491(klassisk/skøyting) dersom vi inntar gjennomsnittshastigheter i alle terrengtyper opp, ned og flatt, som forklaringsvariable for den avhengige variabelen gjennomsnittshastighet på hele løpet. Dette er ikke overraskende da oppover er det segmentet som alene best korrelerer med gjennomsnittshastigheten på hele løpet. De øvrige semipartielle korrelasjonene er i klassisk/skøyting 0.145/0.114 og 0.051/0.035 for respektive ned- og flatseksjonen. Vi har sett at i skøyting korrelerer flatt best med gjennomsnittshastighet på hele løpet sammenlignet med nedover og i klassisk er korrelasjonene nær like. Den semipartielle korrelasjonen er større for nedoverseksjonen versus flatseksjonen ikke fordi nedoverseksjonen best korrelerer med gjennomsnittshastighet på hele løpet smln. med flatseksjonen, men fordi flatseksjonen i større grad enn nedoverseksjonen korrelerer med oppoverseksjonen. Flatpartiet gir derfor mindre tilleggsinformasjon til den avhengige variabelen enn det nedoverpar-

tiet gjør. Å hevde at en høy hastighet i nedoverseksjonen er viktigere for sluttresultatet enn en høy hastighet i flatseksjonen, som følge av en høyere semipartiell korrelasjon er altså positivt feil.

Konklusjon:
Det var, ikke overraskende, tydelige sammenhenger mellom hastighet i de ulike seksjoner og gjennomsnittshastighet på hele løpet. Det var likevel bare oppoverseksjonen, som ga signifikante sammenhenger etter alle metoder i begge disipliner.
Med tanke på at man tilbringer langt mer tid i oppoverseksjonen enn i flatseksjonen og at man tilbringer lengre tid i flatseksjonen enn i nedoverseksjonen forsterker dette antydningene ovenfor om at oppoverseksjonen er viktigst for sluttresulatet deretter flatseksjonen og nedoverseksjonen tilslutt.
Det er ingen klare tegn, jfr. analysene ovenfor, på at løperne i forhold til sluttresultat pacer spesielt ulikt i de ulike terrengtypene. Det kan likevel være interessant, i nye undersøkelser, å studere hvorvidt de beste ikke er best i de letteste partiene av løypa og dermed i større grad klarer å bruke disse partiene til restiusjon sammenlignet med sine svakere konkurrenter.
Når jeg splittet på rundene fant jeg lite avvik fra om vi så hele løpet under et.

6.4.5 Puls i de ulike segmenter og gjennomsnittspuls mot gjennomsnittshastighet totalt(H_{10}) - 2014-data jfr. tabell 20 på neste side

Normalfordelte data:
Data er, i henhold til en shapiro wilk test, ikke normalfordelt i noe terreng verken for hele løpet, rundene, eller den enkelte runde.
Aktuelle metoder er 3,4,5,7 og 8.
Korrelasjonsanalyse:
Ved plotting av dataene observerer jeg en uteligger(klart lavest puls, men også klart høyest hastighet). Uteliggeren påvirker ikke retningen på Pearson's r, men medfører at vi går fra klart signifikante sammenhenger til ikke-signifikante, eller såvidt signifikante sammenhenger. Kendall's t og Spearman's r_s er, som vi vet, robust mot uteliggere og jeg velger i analysen her kun å fokusere på metodene 7 og 8.
Vi konstaterer at jo fortere du går jo lavere gjennomsnittspuls har man. I klassisk gir M7 og M8 signifikante sammenhenger for alle terrengtyper bortsett fra nedover ved bruk av M7. I skøyting totalt og opp er korrelasjonene identiske med klassiskdisiplinen etter M7 og M8. I flatseksjonen og nedoverseksjonen har vi høyere korrelasjoner etter M7 og M8 enn i klassisk og slik

Tabell 20: Korrelasjon mellom gjennomsnittshastighet totalt og gjennomsnittspuls totalt og puls i de enkelte segmenter opp, flatt og ned - 2014

Metode M n=6	Observert korrelasjon klassisk/skøyting	P-verdi klassisk/skøyting	Konfidensintervall(0.95) klassisk/skøyting
M1 totalt	-0.72/-0.79	0.06/0.03	(-0.97,0.1)/(-0.98,-0.11)
opp	-0.73/-0.82	0.05/0.02	(-0.97,0.06)/(-0.98,-0.19)
flatt	-0.7/-0.77	0.06/0.04	(-0.97,0.14)/(-0.98,-0.04)
ned	-0.67/-0.73	0.07/0.05	(-0.97,0.19)/(-0.97,0.07)
M2 totalt	**	0.05/0.03	(-0.96,0.19)/(-0.97,0.01)
opp	**	0.05/0.02	(-0.97,0.16)/(-0.98,-0.06)
flatt	**	0.06/0.03	(-0.96,0.22)/(-0.97,0.07)
ned	**	0.07/0.05	(-0.96,0.27)/(-0.96,0.17)
M3 totalt	**	0.03/0.004	
opp	**	0.03/0.004	
flatt	**	0.04/0.006	
ned	**	0.06/0.003	
M4' totalt	**	0.06/0.03	(-1,0.26)/(-1,0.05)
opp	**	0.05/0.02	(-1,0.22)/(-1,-0.02)
flatt	**	0.06/0.04	(-1,0.3)/(-1,0.12)
ned	**	0.07/0.05	(-1,0.36)/(-1,0.23)
M7 totalt	-0.87/-0.87	0.008/0.008	
opp	-0.87/-0.87	0.008/0.008	
flatt	-0.73/-0.83	0.03/0.008	
ned	-0.6/-0.97	0.07/0.001	
M8 totalt	-0.94/-0.94	0.008/0.008	
opp	-0.94/-0.94	0.008/0.008	
flatt	-0.89/-0.93	0.02/0.008	
ned	-0.83/-0.99	0.03/0.001	
M5"	Korrelasjon	Konf.int. norm.app.	Konf.int. bootstrap
totalt	-0.72/-0.79	(-1,-0.42)/(-1,-0.53)	(-0.96,-0.43)/(-0.96,-0.58)
M6"'	Korrelasjon	Konf.int. norm.app.	Konf.int. bootstrap
totalt	-0.72/-0.79	(-1,-0.17)/(-1,-0.34)	(-1,-0.46)/(-1,-0.6)
M4'	Skjevhet: g1x, g1y	Kurtosis: g2x, g2y	Momenter: E(r^3), E(r^4)
totalt	-1.08,0.06/-1.11,0.04	0.54,-1.43/0.63,-1.64	-0.002,0.08/-0.001,0.08
opp	-1.05,0.06/-1,0.04	0.5,-1.43/0.37,-1.64	-0.002,0.08/-0.001,0.08
flatt	-1.06,0.06/-1.18,0.04	0.48,-1.43/0.77,-1.64	-0.002,0.08/-0.001,0.08
ned	-1.07,0.06/-1.26,0.04	0.49,-1.43/0.97,-1.64	-0.002,0.08/-0.001,0.08
M5"	Skjevhet $\hat{\rho}$	Standardavvik $\hat{\rho}$	
totalt	-0.1/-0.08	0.15/0.14	
M6"'	Skjevhet $\hat{\rho}$	Standardavvik $\hat{\rho}$	
totalt	0.04/0.04	0.28/0.23	

**Identisk med M1.

at begge er klart signifikante.

Korrelasjoner mellom de ulike segmentene - pacing:

Det er svært høy korrelasjon, i begge teknikker, mellom to og to av de fire variablene gjennomsnittspuls på hele løpet, gjennomsnittspuls oppover, gjennomsnittspuls bortover og gjennomsnittspuls nedover. I tillegg er det for alle seksjoner en høy negativ korrelasjon med gjennomsnittshastighet på hele løpet. Det er derfor lite, som tyder på at utøverne m.h.p. puls pacer ulikt i de ulike seksjoner. Dette innebærer at rangeringen av løperne m.h.p. puls bør være nokså like enten vi studerer hele løpet under et, eller et og et segment for hver av de to teknikkene klassisk og fristil. Rangeringene med 1 som dårligst og 6 som best på løperne (A,B,C,D,E,F) var i henholdsvis klassisk/skøyting for hastighet totalt(1,2,3,5,4,6/1,2,3,5,6,4) og puls totalt(4,6,5,3,2,1/4,5,6,3,2,1), puls opp(4,6,5,3,2,1/4,5,6,3,2,1), puls ned(4,6,3,5,1,2/4,6,5,3,2,1) og puls bort(4,6,3,5,2,1/4,5,6,2,3,1), som bekrefter vår antagelse.

Konklusjon:

Det er en klar negativ sammenheng mellom gjennomsnittshastighet på hele løpet og puls, men det er ingenting i data, som tyder på at dersom man har lavere puls i f.eks bortover og nedoverseksjonen så har man høyere puls i oppoverseksjonen. Ved å studere runde 1 og runde 2 hver for seg er det i klassisk tegn til at man oppover kjører hardere pulsmessig i runde 1 kontra runde 2.

6.4.6 Hvor hardt du tar i av gjennomsnittspuls i de ulike segmentene mot gjennomsnittelig hastighet totalt(H_{11}) - 2014-data

Det var ingen signifikante sammenhenger verken i oppover-, bortover, eller nedoversegmentet i skøyting eller klassisk. Jeg fant dog en tendens mot negativ korrelasjon i oppoversegmentet både i klassisk og skøyting - m.a.o. jo hardere du må ta i av gjennomsnittspulsen i oppoversegmentet jo dårligere blir gjennomsnittshastigheten. I bortoverseksjonen i klassisk var det en tendens til positiv korrelasjon slik at jo hardere man klarer å ta i av gjennomsnittspulsen her jo høyere blir gjennomsnittshastigheten. I skøyting fant jeg ingen tendens. I nedoverseksjonen, i både klassisk og skøyting, var det en tendens til positiv korrelasjon og tendensen var tydeligst i skøyting.

Konklusjon:

Vi har tidligere, jfr. 6.4.4 på side 100, sett at oppoverseksjonen er viktigst for sluttresultatet og at jo fortere du går der jo bedre blir sluttresultatet.

Samtidig ser vi nå, noe overraskende, at de som går raskest er de som presser pulsen minst i oppoverseksjonen. Dette tyder på at de beste må gå arbeidsøkonomisk og teknisk bedre enn de svakere oppover.
Vi har tidligere sett i de lettere partiene, spesielt i nedoverseksjonen, at hastigheten der er en anelse mindre korrelert med gjennomsnittshastighet på hele løpet enn hastigheten i oppoverseksjonen. Denne observasjonen brukte jeg som et mulig argument for at de beste var flinkere til å hvile nedover. Det er derfor overraskende at jo hardere du tar i av gjennomsnittpulsen nedover jo bedre blir gjennomsnittshastigheten! Det krever en større undersøklse til, med tilgang på langt mere data, for å komme til bunns i dette.

6.4.7 Gjennomsnittshastighet på hele løpet sammenlignet med hastighet i de ulike segmenter (H_{12}) - 2013-data jfr. tabell 21 på neste side

Normalfordelte data:
Jeg fikk lave p-verdier når jeg kontrollerte(shapiro wilk test) for hvorvidt det forelå en bivariat normalfordeling. I klasisk/skøyting, for hastighet for henholdsvis opp, ned og flatt, i fordeling med gjennomsnittshastighet på hele løpet fikk jeg p-verdiene 0.04/0.05, 0.006/0.01, 0.06/0.15. Tilsvarende fikk jeg for runde 1 isolert p-verdiene 0.11/0.02, 0.08/0.03, 0.14/0.001 og runde 2 isolert p-verdiene 0.006/0.66, 0.33/0.08, 0.21/0.05.
Jeg finner lite grunnlag for å anta bivariat normalfordeling i motsetning til hva vi gjorde for 2014-data.
Aktuelle metoder blir derfor 3,4,5,7 og 8.
Korrelasjonsanalyse:
I klassisk gir alle aktuelle metoder for alle terrengtyper signifikante, eller svært nær signifikante resultater. I skøyting har vi signifikante resultater for alle metoder oppover, i flatt er det bare M8, som ikke gir signifikans, eller svært nær signifikante resultater. Det bemerkes at M5 ikke lot seg beregne for flatt skøyting grunnet to observasjoner med like hastigheter i flatpartiet i skøyting og kun 5 observasjoner. Resultatene vi får i flatpartiet og oppoverpartiet er nokså tilsvarende de resultater vi fikk i 2014.
I skøyting nedover har vi korrelasjonene 0 og 0.1 for de respektive metodene M7 og M8, observert Pearson's r=0.32 og konfidensintervallene (-0.83,0.82), (-1,1) og (-0.35,1) for henholdsvis M3, M4 og M5. Det er m.a.o. vanskelig å hevde at det foreligger noen positiv tendens mellom hastighet nedover i skøyting og gjennnomsnittshastighet på løpet. Korrelasjonene her er langt lavere enn de tilsvarende usignifikante korrelasjonene vi fikk i 2014 og kan i større grad enn i 2014 tyde på en pacingstrategi.
Korrelasjoner mellom de ulike segmentene - pacing:

Tabell 21: Korrelasjon mellom gjennomsnitthastighet totalt og gjennomsnittshastighet i de ulike segmenter opp, flatt og ned - 2013

Metode M n=5	Observert korrelasjon klassisk/skøyting	P-verdi klassisk/skøyting	Konfidensintervall(0.95) klassisk/skøyting
M1			
opp	0.98/0.99	0.002/0.0002	(0.86,0.999)/(0.96,0.999)
flatt	0.91/0.93	0.02/0.01	(0.41,0.995)/(0.52,0.996)
ned	0.89/0.32	0.02/0.3	(0.34,0.99)/(-0.75,0.94)
M2			
opp	**	0.0002/0.000003	(0.77,0.999)/(0.93,0.999)
flatt	**	0.01/0.01	(0.22,0.99)/(0.33,0.99)
ned	**	0.01/0.3	(0.15,0.99)/(-0.74,0.93)
M3			
opp	**	0.02/0.008	
flatt	**	0.02/0.03	
ned	**	0.008/0.34	
M4'			
opp	**	0.002/0.0002	(0.62,1)/(0.81,1)
flatt	**	0.02/0.01	(0.13,1)/(0.24,1)
ned	**	0.02/0.3	(0.07,1)/(-1,1)
M7			
opp	0.8/1	0.04/0.008	
flatt***	0.8/0.74(0.70)	0.04/0.04	
ned	1/0	0.008/0.59	
M8			
opp	0.9/1	0.04/0.008	
flatt****	0.9/0.82	0.04/0.07	
ned	1/0.1	0.008/0.475	
M5"	Korrelasjon	Konf.int. norm.app.	Konf.int. bootstrap
opp	0.98/0.99	(0.66,1)/(0.987,1)	(0.96,1)/(0.989,1)
flatt	0.91/0.93	(0.58,1)/(0.85,1)	(0.81,0.97)/x
ned	0.89/0.32	(0.7,1)/(-1,1)	(0.79,1)/(-0.35,1)
M6"'	Korrelasjon	Konf.int. norm.app.	Konf.int. bootstrap
opp	0.98/0.99	(0.88,1.08)/(0.96,1)	(0.96,1)/(0.99,1)
flatt	0.91/0.93	(0.58,1.23)/(0.65,1)	(0.82,1)/(0.86,1)
ned	0.89/0.32	(0.54,1.25)/(-0.6,1)	(0.79,1)/(-0.29,1)
M4'	Skjevhet: g1x, g1y	Kurtosis: g2x, g2y	Momenter: E(r^3), E(r^4)
opp	0.7,0.41/0.48,0.64	-0.35,-0.88/-0.83,-0.59	0.01,0.13/0.01.0.13
flatt	-0.17,0.41/0.34,0.64	-1.7,-0.88/-0.64,-0.59	-0.003,0.13/0,01,0.13
ned	0.02,0.41/0.01,0.64	-1.72,-0.88/-1.56,-0.59	0.0003,0.13/0.0004,0.13

**Identisk med M1.

***$t_b(t_a)$ ved ties og p-verdi beregnes da ved normaltilnærming med kontinuitetskorreksjon.

****r_{sb} ved ties og p-verdi beregnes da ved normaltilnærming med kontinuitetskorreksjon.

Det er klart mindre positiv korrelasjon mellom hastighet i nedoverterrenget og gjennomsnittelig hastighet på hele løpet, sammenlignet med oppover- og bortoverseksjonen, i skøyting. Når jeg studerere korrelasjonene mellom de ulike terrengtypene i skøyting konstaterer vi, ikke overraskende, at opp og flatt er høyt korrelert med hverandre og at ned er lavt korrelert med opp og flatt(i klassisk er alle seksjoner høyt korrelert med hverandre). Dette tyder på at utøverne m.h.p. hastighet i skøyting pacer ulikt i nedoverseksjonen sammenlignet med de to andre seksjonene oppover og bortover. Vi bør av den grunn, i skøyting, oppdage små forskjeller mellom rangeringene av løperne relatert til hastighet i oppover og bortover når vi sammenligner med rangeringen til slutt, men derimot bør rangeringen i nedoverseksjonen være mer tilfeldig sammenlignet med sluttrangeringen. I klassisk bør det være høyt samsvar mellom rangeringen i de enkelte seksjoner og sluttrangeringen. Rangeringene med 1 som dårligst og 5 som best på løperne (A,B,C,D,E) var i henholdsvis klassisk/skøyting totalt(4,3,2,1,5/2,4,3,1,5), opp(4,3,1,2,5/2,4,3,1,5), ned(4,3,2,1,5/1,3,2,4,5) og bort(3,4,2,1,5/4,2,3,1,5), som bekrefter våre antagelser, men selv i skøyting var likevel den beste best i alle seksjoner - også nedover!

Semipartielle korrelasjoner - pacing:

Ved å bruke lineær multippel regresjon herunder studere de semipartielle korrelasjonene konstaterer jeg at oppoverseksjonen er den langt viktigste variabelen for å forklare og predikere gjennomsnittelig hastighet på hele løpet(avhengig variabel) dersom vi inntar hastigheter i alle terrengtyper opp, ned og flatt, som forklaringsvariable. Den semipartielle korrelasjonen for oppover er på 0.412/0.35(klassisk/skøyting) og høyere enn for nedover og bortover - dette er ikke overraskende da oppover er det segmentet som alene best korrelerer med gjennomsnittshastigheten. De øvrige semipartielle korrelasjonene er i klassisk/skøyting 0.047/0.069 og 0.048/0.079 for respektive ned- og flatseksjonen. Jeg har videre kommentert at i skøyting korrelerer flatt best med gjennomsnittshastighet på hele løpet, sammenlignet med nedover, og i klassisk er korrelasjonene nær like. I skøyting kunne man derfor forventet at den semipartielle korrelasjonen for nedoverseksjonen skulle være høyere enn for flatterrenget grunnet flatseksjonens høye korrelasjon med oppoverseksjonen sammenlignet med nedoverseksjonen. Men siden nedoverseksjonen i seg selv har svært liten forklaringseffekt på gjennomsnittshastigheten på hele løpet i skøyting, grunnet den lave korrelasjonen med gjennomsnittshastigheten på hele løpet, blir den semipartielle korrelasjonen for nedoverseksjonen også svært lav.

I klassisk er det ikke overraskende at de semipartielle korrelasjonene for henholdsvis bortover- og nedoverseksjonen er små og like - både nedover og bortover er nær likt og høyt korrelert med oppoverseksjonen i tillegg til at

begge er høyt og likt korrelert med gjennomsnittshastigheten.
Rundene isolert:
I runde 1 har vi tilsvarende resultater i oppoverseksjonen, som for hele løpet sett under et, i begge skidisipliner. I flatseksjonen får vi noe lavere sammenheng ved bruk av M7 og M8 i skøyting og i nedoverseksjonen klart høyere positiv korrelasjon i skøyting. I runde 2 er det noe mer sammenheng i klassisk oppover etter metode 7 og 8 enn det var når vi så hele løpet under et. Bortover er det lavere sammenheng i runde 2 i klassisk enn løpet sett under et. I nedoverseksjonen er det, i skøyting, verd å merke seg en negativ tendens mot den klare positive sammenhengen i runde 1.

Konklusjon:
Det var, ikke overraskende, tydelige sammenhenger mellom hastighet i oppover- og flatseksjonen og gjennomsnittshastighet på hele løpet. Det var likevel bare oppoverseksjonen, som ga signifikante sammenhenger etter de aktuelle metodene M3, M4, M7 og M8 i begge disipliner, dog tilsvarende resultater i flatseksjonen. I klassisk nedover var det tilsvarende signifikante resultater, som oppover. I skøyting er det ikke grunnlag for å hevde noen som helst sammenheng mellom hastighet nedover og gjennomsnittelig hastighet på hele løpet.
Med tanke på at man tilbringer langt mer tid i oppoverseksjonen enn i flatseksjonen, og at man tilbringer lengre tid i flatseksjonen enn i nedoverseksjonen, forsterker dette antydningene ovenfor om at oppoverseksjonen er viktigst for sluttresulatet, deretter flatseksjonen og nedoverseksjonen minst viktigt(og ikke viktig i det hele tatt i skøyting).
Det er lite holdepunkter for å hevde at det er store forskjeller på hvordan løperne pacer på de ulike rundene. Konklusjonene samsvarer i stor grad med konklusjonene vi hadde ved tilsvarende undersøkelser på 2014-dataene.

6.4.8 Puls i de ulike segmenter og gjennomsnittspuls mot gjennomsnittelig hastighet(H_{13}) - 2013-data

Konklusjon:
Jeg finner ingen signifikante sammenhenger mellom puls og gjennomsnittshastighet totalt eller for de enkelte seksjoner. Jeg finner likevel en negativ tendens i klassisk og enda klarere i skøyting. Resultatene blir bekreftet i de enkelte seksjoner. Når vi skiller på rundene ser vi at i skøyting(totalt og i hver enkelt seksjon) er den negative tendensen i skøyting noe mindre negativ i runde 1 enn når man ser på begge rundene under et og noe mer negativ i runde 2 enn når man ser på begge rundene under et. I klassisk er det lite forskjeller på rundene.

Det er, sammenlignet med 2014-data, en svakere tendens i 2013, men retningen på den lineære sammenhengen og monotonisiteten er likevel den samme.

6.4.9 Hvor hardt du tar i av gjennomsnittspuls i de ulike segmentene sammmenlignet med gjennomsnittelig hastighet(H_{14}) - 2013-data

Konklusjon:
Jeg finner ingen signifikante sammenhenger i noen av seksjonene opp, flatt, eller ned, verken i klassisk eller i skøyting. Det er dog en negativ tendens i oppoverseksjonen og denne er tydeligst i skøyting. Jo hardere du tar i av gjennomsnittspuls jo lavere blir gjennomsnittshastigheten. I flatterrenget har vi en anelse positiv tendens, men slik at i skøytedisiplinen er denne nær null. I nedoverterrenget har vi en bitteliten positiv tendens i klassisk og en liten negativ tendens i skøyting. Dette innebærer de samme konklusjoner vi fikk for 2014 bortsett fra den lille negative tendensen nedover i skøyting.

6.4.10 Korrelasjoner fra to uavhengige utvalg 2013/2014

Data ga ikke grunnlag for å benytte noen av testene i 2.7.1 på side 29. Korrelasjonene mellom puls og gjennomsnittshastighet for kvinner i henholdsvis 2013 og 2014 var klart forskjellige, og det ville vært interessant å teste om det var signifikante forskjeller, dersom vi hadde hatt tilgang på en relevant test. Igjen er lite data et stort drawback ved statistisk analyse.

6.4.11 Partielle korrelasjoner - 2014

I seksjonen her studerer vi hvorvidt sammenhengen/korrelasjonen mellom gjennomsnittelig hastighet i skøyting for kvinner på hele løpet i 2014 og henholdsvis VO2max i diagonalgang og VO2max i staking er misledende dersom vi kontrollerer for antall treningstimer nedlagt(mai til oktober 2014) jfr. diskusjonen i 5.1 på side 79. Det ville vært mer realistisk å se på et større tidsrom for nedlagt treningsmengde, men eksempelet illustrerer under enhver omstendighet praktisk anvendelse av partielle korrelasjoner, selvom det ikke kan legges for mye vekt på konklusjonene.

VO2max i diagonalgang(H_{15}): Resultatene fremkommer av tabellen 22 på neste side.
Den multivariate normalantagelsen , som kreves oppfylt ved bruk av Pearsontesten, er ikke oppfylt - en multivariat shapiro wilk test forkastet antagelsen med en p-verdi på 0.00006.

Tabell 22: Partiell korrelasjon: Gjennomsnittelig hastighet i skøyting mot VO2max i diagonalgang kontrollert for treningsmengde - 2014

Utvalgsstørrelse: n=6		
Test	Korrelasjon(pverdi)	Partiell korrelasjon (pverdi*)
Kendall**	0.73(0.028)	-0.07(0.86)
Spearman	0.89(0.008)	-0.03(0.96)
Pearson***	0.88(0.01)	-0.02(0.98)

*Normalapproksimasjon

**Eksakte kritiske grenser for partiellkorrelasjonen er ±0.667 jfr. 5.1.2 på side 83

***Den multivariate normalantagelsen holder ikke: p-verdi=0.00006

Vi konstaterer videre at den opprinnelige korrelasjonen mellom gjennomsnittelig hastighet i skøyting på hele løpet i 2014 og VO2max i diagonalgang for henholdsvis Kendall's t og Spearman's r_s(henholdsvis 0.73 og 0.89) er signifikant forskjellig fra 0. Dersom vi beregner den partielle korrelasjonen mellom gjennomsnittelig hastighet i skøyting på hele løpet i 2014 og VO2max i diagonalgang når vi kontrollerer for antall nedlagte treningstimer ser vi at den opprinnelige signifikante positive sammenhengen er misvisende da eksakte kritiske grenser etter Kendall-testen for å forkaste H_{15} er ±0.667 med en beregnet partiellkorrelasjon $t_{xy.z}$ = -0.07.

VO2max i staking(H_{16}): Resultatene fremkommer av tabellen 23 på neste side.
Den multivariate normalantagelsen , som kreves oppfylt ved bruk av Pearsontesten, er ikke oppfylt - en multivariat shapiro wilk test forkastet antagelsen med en p-verdi på 0.005.
Vi konstaterer videre at den opprinnelige korrelasjonen mellom gjennomsnittelig hastighet i skøyting på hele løpet i 2014 og VO2max i staking verken for Kendall's t, eller Spearman's r_s er signifikant forskjellig fra 0, men det er dog en klar tendens til positiv sammenheng(henholdsvis 0.46 og 0.71). Dersom vi beregner den partielle korrelasjonen mellom gjennomsnittelig hastighet i skøyting på hele løpet i 2014 og VO2max i staking når vi kontrollerer for antall nedlagte treningstimer ser vi tydelig, at den opprinnelige tendensen til positiv sammenheng er misvisende da eksakte kritiske grenser etter Kendalltesten for å forkaste H_{16} er ±0.667 med beregnet partiellkorrelasjon på $t_{xy.z}$ = -0.13.

Tabell 23: Partiell korrelasjon: Gjennomsnittelig hastighet i skøyting mot VO2max i staking kontrollert for treningsmengde - 2014

Utvalgsstørrelse: n=6		
Test	Korrelasjon(pverdi)	Partiell korrelasjon (pverdi*)
Kendall**	0.46(0.14)	-0.13(0.74)
Spearman	0.71(0.051)	-0.06(0.91)
Pearson***	0.71(0.057)	0.18(0.91)

*Normalapproksimasjon

**Eksakte kritiske grenser for partiellkorrelasjonen er ±0.667 jfr. 5.1.2 på side 83

***Den multivariate normalantagelsen holder ikke: p-verdi=0.005

7 Sluttkommentarer

Det er mange fallgruver når man skal interpretere en beregnet korrelasjon mellom to variable X og Y. Først av alt er det viktig å ha klart for seg hva den beregnede korrelasjonen gir uttrykk for og, som vi har sett tidligere, gir Pearson's r et uttrykk for den lineære sammenhengen mellom to variable mens f.eks. Kendall's t og Spearman's r_s gir uttrykk for en monotonistisk sammenheng mellom to variable. Dersom det er andre sammenhenger, enn en lineær eller monotonistisk sammenheng, mellom to variable man ønsker målt, vil nevnte korrelasjonsmål ikke være riktig verktøy.
Dersom vi ønsker å måle hvorvidt det er en lineær, eller monotonistisk sammenheng, mellom to variable X og Y, vil en signifikant korrelasjon ikke bekrefte årsakssammenheng mellom de to variablene. En beregnet signifikant korrelasjon kan fint oppstå ved tilfeldigheter, eller fordi man har oversett en, eller flere andre variable, som påvirker en eller begge variablene X og Y. Det er m.a.o. svært viktig å ha klart for seg om det foreligger en kausal sammenheng mellom variablene X og Y, og om de nevnte variablene igjen står i kausale forbindelser med andre variable. Vi har sett at beregning av partielle og semipartielle korrelasjoner er relevante verktøy å benytte når man studerer kausale forbindelser mellom variable. Først når man har kontroll på de kausale forbindelser vil det være mulig å interpretere beregnede korrelasjoner på en god måte, og derigjennom hindre at korrelasjonen man oppgir i for stor grad er misvisende, eller illusorisk. Det er utallige eksempler fra media på misvisende bruk av korrelasjoner, og forklaringen ligger ofte i at undersøkelsene, som er foretatt, ikke har tatt høyde for hvorvidt det forelig-

ger en kausal forbindelse mellom de aktuelle størrrelsene det oppgis å være en sammenheng mellom.
Forutsatt at vi er bevisst korrelasjonsmålet vi benytter, og kontrollerer aktuelle kausale forbindelser, er det fortsatt stor usikkerhet knyttet til korrelasjoner beregnet på lite data. Inferens, knyttet til korrelasjonskoeffisientene Pearson's r ved trekk fra den bivariate normalfordeling, Kendall's t og r_s, er, som vi har sett tidligere, forbundet med store standardfeil. Uansett, hva de nevnte korrelasjonskoeffisientene er, vil standardfeilene til Pearson's r ved trekk fra den bivariate normalfordeling, Kendall's t og r_s være henholdsvis av orden $(1-\rho^2)/\sqrt{n}$, $\sqrt{2/n}$ og $\sqrt{3/n}$. Det er altså generelt vanskelig å lokalisere populasjonskorrelasjonen veldig nøyaktig med mindre n ligger i intervallet 30-40 eller høyere. Poenget er at man skal være svært forsiktig med og legge for mye i korrelasjonskoeffisienter beregnet fra data hvor n er lav med mindre vi har mange tilgjengelige målinger.
Det lar seg gjøre å studere sammenhenger mellom variable uten å benytte korrelasjon. Et svært vanlig verktøy er regresjon der vi opererer med en avhengig variabel Y og en eller flere uavhengige variable $X_1, X_2, ..., X_n$. Ved bruk av regresjon er det naturlig at de uavhengige variablene forklarer den avhengige variabelen slik at det er årsakssammenheng(funksjonell sammenheng)mellom hver enkelt uavhengige variabel og den avhengige variabelen. Ved bruk av korrelasjon er hverken X eller Y uavhengig variabel. Regresjon kalles å tilpasse en linje/kurve, plan/flate, eller hyperplan/hyperflate til dataene og kan f.eks. brukes til å forutsi/predikere en variabelverdi ut fra de uavhengige variablene. Ved å dekomponere totalvariasjonen $\sum_i (Y_i - \bar{Y})^2 (SS_{total})$ til Y rundt sitt gjennomnitt får vi:

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{Y})^2 \tag{137}$$

$$\Updownarrow$$

$$1 = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{\sum_i (Y_i - \bar{Y})^2} + \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2} \tag{138}$$

, der $\hat{Y}_i$ er de tilpassede verdiene fra regresjonsanalysen, $\bar{Y}_i = \sum_i (Y_i/n)$ og n er antall observasjoner. Vi konstaterer at $\sum_i (\hat{Y}_i - \bar{Y})^2 (SS_{regresjon})$ forklarer variansen i Y forklart av de tilpassede verdiene via $X_1, X_2, ..., X_n$ og at $\sum_i (Y_i - \hat{Y}_i)^2 (SS_{feil})$ er den andelen av variasjonen i Y, som ikke kan forklares av $X_1, X_2, ..., X_n$. Den multiple korrelasjonskoeffisienten $R^2 = \frac{SS_{regresjon}}{SS_{total}} = 1 - \frac{SS_{feil}}{SS_{total}}$ vil da være et mål for hvor godt modellen(kurven/flaten/hyperflaten) er tilpasset de observerte data.
I mitt videre arbeid med korrelasjoner og regresjon vil jeg ha stort fokus på

kausale forbindelser mellom aktuelle variable.

Til slutt ønsker jeg å takke min veileder Anders Rygh Swensen, professor i statistikk og biostatistikk ved Matematisk institutt ved Universitetet i Oslo, for å ha gitt meg hint og råd i forbindelse med oppgaven, og for alltid å ha vært positiv til mitt arbeide og engasjement.
Jeg ønsker også å takke Olympiatoppen og Norges Skiforbund langrenn representert ved Øyvind Sandbakk, som har gitt meg tilgang på data på noen av Norges beste langrennsløpere til bruk i oppgaven. Jeg ønsker å takke Øyvind Sandbakk og Conor M. Bolger for å ha tatt meg i mot ved Olympiatoppens lokaler på Toppidrettssenteret i Granåsen der jeg ble vist rundt på laben herunder 'utsatt' for en rekke vanlige tester. Under besøket i Trondheim ble jeg også satt bedre inn i samarbeidsprosjektet Olympiatoppen hadde med Norges Skiforbund, se seksjon 1.1 på side 6, hvor jeg fikk bidra med statistiske undersøkelser.
Til slutt vil jeg takke Ann Magdalen Hegge og Jan Kocbach, som tok meg i mot på Beitostølen, i forbindelse med den nasjonale åpningen av langrennssesongen i November 2014. Ann og Jan ga meg innsikt i hvordan GPS-dataene jeg har brukt i oppgaven ble innsamlet og hvordan dataene umiddelbart etter innsamlingen ble lagt frem for utøvere og trenere.

Figurer

1 Chart1 fra [12] 32
2 Utdrag fra Chart 1 33
3 Misledende korrelasjon 1 80
4 Misledende korrelasjon 2 81

Tabeller

1 Korrelasjonsindekser 9
2 Funksjoner brukt i R 13
3 Kontingenstabell for to binære variable 1 37
4 Glidertest - korrelasjonskoeffisienten η 38
5 Glidertest - anova 39
6 Rate of personal effort i klassisk og skøyting - kontinuerlig skala. 55
7 Rate of personal effort i klassisk og skøyting - ordinal skala 1 . 55
8 Rate of personal effort i skøyting og klassisk - ordinal skala 2 . 57
9 VO2max for kvinner og menn - ordinal og nominal skala . . . 62
10 VO2max for kvinner og menn - ordinal skala 62
11 Kontingenstabell for to binære variable 2 63
12 Frekvensfordelingen til S = P - Q 65
13 Eksakte og approksimerte kritiske verdier for Spearman's r_s . 71
14 Fordelingen til P(overensstemmende par) gitt N=9 og n=3 for rangeringene A og B 72
15 Fremgangsmåte for å finne 1.ordens partiellkorrelasjon mellom X og Y 82
16 Korrelasjon mellom høyde og gjennomsnittshastighet totalt og i de enkelte segmentene opp, flatt og ned - 2014 94
17 Korrelasjon mellom VO2max diagonalgang og gjennomsnittshastighet totalt og i de enkelte segmenter opp, flatt og ned - 2014 96
18 Korrelasjon mellom VO2max staking og gjennnomsnittshastighet totalt og i de enkelte segmenter opp, flatt og ned - 2014 97
19 Korrelasjon mellom gjennomsnitthastighet totalt og gjennomsnittshastighet i de ulike segmenter opp, flatt og ned - 2014 . . 101
20 Korrelasjon mellom gjennomsnittshastighet totalt og gjennomsnittspuls totalt og puls i de enkelte segmenter opp, flatt og ned - 2014 104
21 Korrelasjon mellom gjennomsnitthastighet totalt og gjennomsnittshastighet i de ulike segmenter opp, flatt og ned - 2013 . . 107

22 Partiell korrelasjon: Gjennomsnittelig hastighet i skøyting mot VO2max i diagonalgang kontrollert for treningsmengde - 2014 111
23 Partiell korrelasjon: Gjennomsnittelig hastighet i skøyting mot VO2max i staking kontrollert for treningsmengde - 2014 112

A Appendiks 1: Metodene

Jeg gir her en kort oversikt over metodene, som blir brukt til å teste hypotesene ovenfor. Teorien bak de statistiske testene, som er implementert i de ulike metodene, er redegjort for i kapitlene 2-5. Metodene er implementert, som funksjoner i R, jfr. vedlegg 'Vedlegg1ProgrammeringsfilMasteroppgaveEinarChristopherWellén.R', og jeg oppgir her argumentene til metodene og hva de returnerer. De fleste av funksjonene, som inngår i metodene er funksjoner innebygd i R jfr. tabell 2 på side 13. Der jeg selv har implementert kode, eller hentet fra andre, blir dette kommentert.
Aller først oppgis funksjonen, som avgjør om data vi benytter kan antas å komme fra univariate, bivariate eller trivariate fordelinger.

- **Normalsjekk - sjekker univariat, bivariat og trivariat normalfordeling**
  `Aktuelle hypoteser:`
  H_0:X(Y, Z) er univariat normalfordelt.
  H_0:(X,Y)((X,Z),(Y,Z)) er bivariat normalfodelt.
  H_0:(X,Y,Z) er trivariat normalfordelt.
  $n \geq 3$
  Vi forkaster hypotesen om normalantagelse dersom p-verdien typisk er mindre enn 0.05.
  `Argumentene til funksjonen er:`
  1)tre likt dimensjonerte vektorer - legg ved den samme vektoren flere ganger hvis du bare trenger å sjekke 1 eller 2 variable.
  `Funksjonen returnerer:`
  1)Pverdi X,Y,Z,XY,XZ,YZ,XYZ
  R: Package(functions): stats(shapiro.test),mvnormtest(mshapiro.test).

- **Metode 1 - Eksakt målefordeling til Pearson's r når data kommer fra en bivariat normalfordeling jfr. kap.2 Pearson's r**
  `Aktuelle hypoteser:` H_0: $\rho = \rho_1$, der ρ_1 kan være alt mellom -1 og 1.
  n$\geq$3
  `Argumentene til funksjonen er:`
  1) To likt dimensjonerte vektorer x og y, som kommer fra en bivariat normalfordeling.
  2) Den antatte populasjonskorrelasjonen $\rho = \rho_1$.
  3) Nedre og øvre prosentandelgrense for konfidensintervallet vi ønsker.
  `Funksjonen returnerer:`
  1) Pearson's beregnet r.

2) P-verdi.
3) Nedre konfidensintervallgrense for ρ for observert r og n.
4) Øvre konfidensintervallgrense for ρ for observert r og n.
R: Package(functions): SuppDists(pPearson,qPearson), stats(cor), base(length).

- **Metode 2 - Tilnærmet normalfordeling av Fisher-transformasjonen av Pearson's r når data kommer fra en bivariat normalfordeling jfr. kap.2 Pearson's r**
  H_0: $\rho = \rho_1$, der ρ_1 kan være alt mellom -1 og 1.
  n$\geq$11
  `Argumentene til funksjonen er:`
  1) To likt dimensjonerte vektorer x og y, som kommer fra en bivariat normalfordeling.
  2) Den antatte populasjonskorrelasjonen $\rho = \rho_1$.
  3) Nedre og øvre prosentandelgrense for konfidensintervallet vi ønsker.
  `Funksjonen returnerer:`
  1) Pearson's beregnet r.
  2) P-verdi.
  3) Nedre konfidensintervallgrense for observert r og n.
  4) Øvre konfidensintervallgrense for observert r og n.
  R: Package(functions): stats(pnorm, qnorm, cor),base(tanh, log, length, sqrt).
  Har selv implementert Fishertransformasjonen og det fulle uttrykket for standardavviket til fishertransformasjonen.

- **Metode 3 - Eksakt permutasjonsfordeling jfr. kap 3 Permutasjonstester og bootstrapping - Pearson's r**
  H_0: $\rho = 0$.
  n$\geq$ 3
  `Argumentene til funksjonen er:`
  1) To likt dimensjonerte vektorer x og y, som kommer fra en ukjent bivariat fordeling.
  2) Nedre og øvre prosentandelgrense for kritiske grenser.
  `Funksjonen returnerer:`
  1) Pearson's beregnet r.
  2) P-verdi.
  3) Nedre kritiske grense for r for de gitte dataene x og y.
  4) Øvre kritiske grense for r for de gitte dataene x og y.

R: Package(functions): stats(cor), base(length, factorial, do.call, rbind, rep, ceiling, floor), combinat(permn).
Har selv implementert et par mindre snutter.

- **Metode 4 - permutasjonstest(t-tilnærming) jfr. kap 3 Permutasjonstester og bootstrapping - Pearson's r**
  H_0: $\rho = 0$.
  $n \geq 3$
  `Argumentene til funksjonen er:`
  1) To likt dimensjonerte vektorer x og y, som kommer fra en ukjent bivariat fordeling.
  2) Nedre og øvre prosentandelgrense for konfidensintervallet vi ønsker.
  `Funksjonen returnerer:`
  1) Pearson's beregnet r.
  2) P-verdi.
  3) Nedre konfidensintervallgrense for r for observert r og n under H_0: $\rho = 0$.
  4) Øvre konfidensintervallgrense for r for observert r og n under H_0: $\rho = 0$.
  5) g1x,g2x,g1y,g2y,Er3,Er4, der resultatene ovenfor kun kan brukes dersom Er3 og Er4 er lave <=> g1x,g2x,g1y,g2y er lave.
  R: Package(functions): stats(cor, pt, qt), base(length, sqrt, mean, sum).
  Har selv implementert uttrykkene for g1x,g2x,g1y,g2y,Er3,Er4 og et par småting til.

- **Metode 5 - ikke-parametrisk bootstrapping jfr. kap 3 Permutasjonstester og bootstrapping - Pearson's r**
  H_0: $\rho = 0$ - vi lager konfidensintervaller og dersom 0 ikke befinner seg i intervallet så tyder det på at vi må forkaste H_0 og jo lenger intervallet er vekk fra 0 jo sterkere blir denne konklusjonen.
  $n \geq 3$
  `Argumentene til funksjonen er:`
  1) To likt dimensjonerte vektorer x og y, som kommer fra en ukjent bivariat fordeling.
  2) Nedre og øvre prosentandelgrense for konfidensintervallgrenser.
  `Funksjonen returnerer:`
  1) Pearson's beregnet r.
  2) Nedre konfidensintervallgrense ved normalapproksimasjon.
  3) Øvre konfidensintervallgrense ved normalapproksimasjon.
  4) Nedre konfidensintervallgrense for standard bootstrapintervall.

5) Øvre konfidensintervallgrense for standard bootstrapintervall.
6) Skjevhet og standardavvik til bootstrapestimatet.
NB: Hvis du trekker like x eller y for alle ved en indekssampling n(typisk når få observasjoner - n liten) vil beregningene ikke gå gjennom fordi vi får standardavvik for x.star eller y.star til å være 0.
R: Package(functions): stats(cor, qnorm, sd), base(length, mean, rep, sample, as.integer, sort).
Har ellers benyttet koden brukt i [45].

- **Metode 6 - parametrisk bootstrapping jfr. kap 3 Permutasjonstester og bootstrapping - Pearson's r**
  H_0: $\rho = 0$ - vi lager konfidensintervaller og dersom 0 ikke befinner seg i intervallet så tyder det på at vi må forkaste H_0 og jo lenger intervallet er vekk fra 0 jo sterkere blir denne konklusjonen.
  $n \geq 3$
  `Argumentene til funksjonen er:`
  1) To likt dimensjonerte vektorer x og y, som vi antar kommer fra en kjent bivariat fordeling (og her implementert kun for normalfordelingen).
  2) Nedre og øvre prosentandelgrense for konfidensintervallgrenser.
  `Funksjonen returnerer:`
  1) Pearson's beregnet r.
  2) Nedre konfidensintervallgrense ved normalapproksimasjon.
  3) Øvre konfidensintervallgrense ved normalapproksimasjon.
  4) Nedre konfidensintervallgrense for standard bootstrapintervall.
  5) Øvre konfidensintervallgrense for standard bootstrapintervall.
  6) Skjevhet og standardavvik til bootstrapestimatet.
  R: Package(functions): stats(cor, rnorm, qnorm, sd), base(length, mean, rep, as.integer, sort).
  Har ellers benyttet koden brukt i [45].

- **Metode 7 - Kendall's tau (τ, τ_b og τ_a) jfr. kap 4 Ikke-parametriske korrelasjonsmål**
  H_0: $\rho = 0$.
  $n \geq 3$
  `Argumentene til funksjonen er:`
  1) To likt dimensjonerte vektorer x og y, som kommer fra en ukjent bivariat fordeling.
  2) Nedre og øvre prosentandelgrense for kritiske grenser.
  `Funksjonen returnerer:`
  1) Kendall's t(evt. t_b og t_a hvis ties, men slik at t_b er beregnet ved

normalapproksimasjon med kontinuitetskorreksjon).
2) Eksakt P-verdi hvis ikke ties(normaltilnærming med kontinuitetskorreksjon hvis ties).
3) Nedre kritiske grense beregnet, som om det ikke var ties.
4) Øvre kritiske grense beregnet, som om det ikke var ties.
5) P-verdi ved normaltilnærming av Kendall's t.
6) P-verdi ved normaltilnærming og kontinuitetskorreksjon av Kendall's t.
7) Nedre **konfidensintervallgrense** for Kendall's t ved normaltilnærming av Kendall's t når vi ser på hele populasjonen. n moderat, N stor.
8) Øvre **konfidensintervallgrense** for Kendall's t ved normaltilnærming av Kendall's t når vi ser på hele populasjonen. n moderat, N stor.
R: Package(functions): stats(pnorm, qnorm), base(length, sqrt), Kendall(Kendall), SuppDists(pKendall, qKendall). n moderat, N stor.
Har selv implementert noen mindre snutter.

- **Metode 8 - Spearman's r jfr. kap 4 Ikke-parametriske korrelasjonsmål**
  H_0: $\rho = 0$.
  $n \geq 3$
  `Argumentene til funksjonen er:`
  1) To likt dimensjonerte vektorer x og y, som kommer fra en ukjent bivariat fordeling.
  2) Nedre og øvre prosentandelgrense for kritiske grenser.
  `Funksjonen returnerer:`
  1) Spearman's beregnet r.
  2) Eksakt p-verdi hvis ikke ties (hvis ties brukes 'AS89'-algoritmen)
  3) Nedre kritiske grense, som om det ikke var ties.
  4) Øvre kritiske grense, som om det ikke var ties.
  5) Ensidig p-verdi ved t-tilnærming av Spearman's r jfr. metode 4 hvis n $\geq$ 11.
  6) Nedre konfidensintervallgrense for Spearman's r ved t-tilnærming under H0 jfr. metode 4 hvis n $\geq$ 11.
  7) Øvre konfidensintervallgrense for Spearman's r ved t-tilnærming under H0 jfr. metode 4 hvis n $\geq$ 11.
  R: Package(functions): stats(pt, qt), base(length, sqrt), pspearman(spearman.test), SuppDists(pSpearman, qSpearmman).
  Har selv implementert noen mindre snutter.

- **Metode 11 - Partielle korrelasjoner jfr. kap 5 Misvisende korrelasjon**
  H_0: pxy.z = 0 - vi forkaster denne dersom p-verdien typisk er lavere enn 0.05.
  $n \geq 4$
  `Argumentene til funksjonen er:`
  1) Tre likt dimensjonerte vektorer x, y og z, som ved beregning av den partielle korrelasjonen.
  `Funksjonen returnerer:`
  1) Partiell korrelasjon for Pearson's r.
  2) Partiell korrelasjon for Kendall's t.
  3) Partiell korrelasjon for Spearman's r.
  4) p-verdi(approksimert) for den partielle korrelasjonen til Pearson's r.
  5) p-verdi(approksimert) for den partielle korrelasjonen til Kendall's t.
  6) p-verdi(approksimert) for den partielle korrelasjonen til Spearman's r.
  R: Package(functions): ppcor(pcor.test).

- **Metode 12 - To uavhengige utvalg under bivariat normalfordeling og Fishertransformasjon jfr. kap.2 Pearson's r**
  H_0: $\rho_1 = \rho_2$
  $n \geq 11$
  `Argumentene til funksjonen er:`
  1) x1,y1,x2,y2, der x1,y1 og x2,y2 kommer fra hver sin bivariate normalfordeling og der x1,y1 er likt dimensjonerte vektorer og x2,y2 er likt dimensjonerte vektorer.
  2) ρ_1 og ρ_2.
  `Funksjonen returnerer:`
  1) Pearson's r1 og r2.
  2) P-verdi.
  3) Nedre konfidensintervallgrense for $\rho_1 - \rho_2$.
  4) Øvre konfidensintervallgrense for $\rho_1 - \rho_2$.
  Inneholder ikke konfidensintervallet null forkaster vi H_0 **R: Package(functions): stats(pnorm, qnorm, cor),base(tanh, log, length, sqrt). Har selv implementert Fishertransformasjonen og det fulle uttrykket for standardavviket til fishertransformasjonen.**

- **Metode 13 - To uavhengige utvalg under bivariat normalfor-**

deling og eksaktfordelingen til Pearson's r jfr. kap.2 Pearson's r
H_0: $\rho_1 = \rho_2$
$n \geq 3$
`Argumentene til funksjonen er:`
1) x1,y1,x2,y2, der x1,y1 og x2,y2 kommer fra hver sin bivariate normalfordeling. x1,y1 er likt dimensjonerte vektorer og x2,y2 er likt dimensjonerte vektorer.
2) n1 og n2, som er antall observasjoner i hver av de to fordelingene.
3) Nedre og øvre prosentandelgrense for konfidensintervallgrenser.
`Funksjonen returnerer:`
1) pa1=øvre grense for ρ gitt r1 og n1 under en konfidensgrad på øvre-nedre.
2) pb1=nedre grense for ρ gitt r1 og n1 under en konfidensgrad på øvre-nedre.
3) pa2=øvre grense for ρ gitt r2 og n2 under en konfidensgrad på øvre-nedre.
4) pb2=nedre grense for ρ gitt r2 og n2 under en konfidensgrad på øvre-nedre.
5) pb2-pa1 og er denne positiv forkaster vi H_0: $\rho_1 = \rho_2$.
6) pb1-pa2 og er denne positiv forkaster vi H_0: $\rho_1 = \rho_2$.
R: Package(functions): stats(cor), SuppDists(pPearson), base(abs). Har selv implementert noe kode.

- **Metode 14 - Semipartielle korrelasjoner jfr. kap 5 Misvisende korrelasjon**
  Bruker funksjonen spcor i R direkte. Vi tester hvorvidt de semipartielle korrelasjonene er null og forkaster hypotesen om at den semipartielle korrelasjonen er null hvis p-verdien typisk er lavere enn 0.05.
  Forklaring metode 14:
  `Argumentene til funksjonen er:`
  1) En matrise x.
  2) Metode('pearson', 'kendall', 'spearman').
  `Funksjonen returnerer:`
  1) De parvis semipartielle korrelasjonene for hvert par av variable gitt andre.
  2) P-verdiene(approksimerte) til hver av de semipartielle korrelasjonene.
  R: Package(functions): ppcor(spcor).

Referanser

[1] Maurice Kendall and Jean Dickinson Gibbons, *Rank Correlation Methods 5.utgave*. Edward Arnold, A division of Hodder & Stoughton, LONDON MELBOURNE AUCKLAND, Femte utgave, 1990

[2] Maurice G. Kendall and Alan Stuart, *The advanced theory of statistics - volume1: Distribution theory*. Charles Griffin & company limited, London, Tredje utgave, 1969

[3] Maurice G. Kendall and Alan Stuart, *The advanced theory of statistics - volume2: Inference and relationship*. Charles Griffin & company limited, London, Tredje utgave, 1973

[4] Maurice G. Kendall and Alan Stuart, *The advanced theory of statistics - volume3: Design and analysis. and time-series*. Charles Griffin & company limited, London, Andre utgave, 1968

[5] Peter Y. Chen and Paula M. Popovich, *Correlation - Parametric and nonparametric measures*. Sage university papers series on quantitative applications in the social sciences, 07-139. Thousand Oaks, CA: Sage., 2002

[6] G.V. Glass and K.D Hopkins, *Statistical methods in education and psychology*. Boston: Allyn & Bacon, Tredje utgave, 1996

[7] W.L. Hays, *Statistics*. New York: Harcourt Brace, Femte utgave, 1994

[8] A.K. Gayen, *The frequency distribution of the product-moment correlation coefficient in random samples of any size drawn from non-normal universes*. Biometrika, 38, 219., 1951

[9] R.A. Fischer, *Frequency-distribution of the values of the correlation coefficient in samples from an indefinitely large population*. Biometrika, 10, 507., 1915

[10] H. Hotelling, *New light on correlation coefficient and its transforms*. J.R. Statist. Soc., B, 15, 193., 1953

[11] W.S.Gosset('Student'), *On the probable error of a correlation coefficient*. Biometrika, 6, 302., 1908

[12] F.N. David, *Tables of the correlation coefficient*. Cambridge Univ. Press, 1938

[13] R.A. Fisher, *On the probable error of a coefficient of correlation deduced from a small sample.* Metron, 1, No. 4, 1., 1921

[14] W. Hoeffding, *The large-sample power of tests based on permutations of observations.* Ann. Math. Statist., 23, 169., 1952

[15] G.E.P. Box, *Non-normality and tests on variances.* Biometrika, 40, 318., 1953

[16] H.E. Daniels, *The relation between measures of correlation in the universe of sample permutations.* Biometrika, 33, 129-35., 1944

[17] H.E. Daniels, *A property of rank correlations.* Biometrika, 35, 416-47., 1948

[18] S.T. David, M.G. Kendall og A.Stuart, *Some questions of distribution in the theory of rank correlation.* Biometrika, 32, 241-52., 1951

[19] H.S. Konijn, *On the power of certain tests for independence in bivariate populations.* The Annals of Mathematical Statistics, 27, 300-23(Correction, Ibid., 29, (1958), 935-6.), 1956.

[20] L.A. Goodman og W.H. Kruskal, *Measures of association for cross classifications.* Journal of the American Statistical Association, 49, 732-64(Correction, Ibid.,52, 578), 1954

[21] L.A. Goodman og W.H. Kruskal, *Measures of association for cross classifications.III:Approximate sampling theory.* Journal of the American Statistical Association, 58, 310-64, 1963

[22] I. Rosenthal, *Distribution of the sample version of the measure of association, Gamma.* J. Amer. Statist. Ass., 59, 460, 1966

[23] G.U. Yule, *On the association of attributes in statistics.* Phil. Trans., A, 194, 257, 1900

[24] G.U. Yule, *On the methods of measuring association between two attributes.* Statist. Soc., 75, 579, 1912

[25] S.E. Fienberg og J.P. Gilbert, *The geometry of a two by two contingency table.* J. Amer. Statist. Ass., 65, 694, 1970

[26] W.G. Cochran, *Some methods for strengthening the common χ^2 test.* Biometrics, 10, 417, 1954

[27] K. Pearson, *On the theory of contingency and its relation to association and normal correlation.* Drapers' Co. Memoirs, Biometric Series, No. 1, London, 1904

[28] H. Cramér, *Mathematical methods of statistics.* Princeton Univ. Press, 1946

[29] S.N. Roy og S.K. Mitra, *An introduction to some non-parametric generalisations of analysis of variance and multivariate analysis.* Biometrika , 43, 361, 1956

[30] G.H. Freeman og J.H. Halton, *Note on an exact treatment of contingency, goodness of fit and other problems of significance.* Biometrika, 38, 141. 1951

[31] K. Pearson, *On the probable error of a coefficient of mean square contingency.* Biometrika, 10, 570, 1915

[32] H.M. Blalock, Jr, *Probabilistic interpretations for the mean square contingency.* J. Amer. Statist. Ass., 53, 102, 1958

[33] P. M. E. Altham, *The measurement of association of rows and columns for an r*s contingency table.* J. R. Statist. Soc., B, 32, 63, 1970

[34] H.A. Simon, *Spurious correlation: A causal interpretation.* American Statistical Association Journal, September, 467-479, 1954

[35] J. Cohen og P. Cohen, *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum, 1.utgave 1975, 2.utgave 1983 og 3.utgave 2003.

[36] R. Rosenthall, *Meta-analytic procedures for social research.* Newbury Park, CA: Sage, 1991

[37] M. Strube, *Averaging correlation coefficients: Influence of heterogenity and set size.* Journal of Applied Psychology, 73, 559-568, 1988

[38] J. B. Caroll, *The nature of the data, or how to choose a correlation coefficient.* Psychometrika, 26, 347-372, 1961

[39] G. V. Glass og K. D. Hopkins, *Statistical methods in education and psychology (3rd ed.).* Boston Allyn og Bacon, 1996

[40] R. J. Wherry SR., *Contributions to correlational analysis.* New York: Academic Press, 1984

[41] J. H. Steiger, *Tests for comparing elements of a correlation matrix.* Psychological Bulletin, 87, 245-251, 1980

[42] B. Efron, *The Jackknife, the Bootstrap, and Other Resampling Plans.* Number 38 in CBMS-NSF Regional Conference Series in Applied Mathematics, Siam, Philadelphia, 1982

[43] J.L. Devore and K.N. Berk, *Modern mathematical statistics with applications.* Duxbury Pr, 2007, ISBN 0534404731.

[44] Michael D. Ernst, *Permutation Methods: A Basis for Exact Inference.* Statistical Science, 2004, Vol. 19, No. 4, 676-685, DOI 10.1214/088342304000000396.

[45] Geir Storvik, *Bootstrapping - Tilleggslitteratur for STK 2120*, Mars 2011.

[46] B.Efron og R. J . Tibshirani, *An introduction to the Bootstrap.* Chapman og Hall, New York, 1993

[47] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and their Applications.* Cambridge Series in Statistics and Probabilistics Mathematics, Cambridge University Press, Cambridge, 1997

[48] R.E. Odeh og J.M.Davenport, *Selected tables in mathematical statistics Volume 10.* Institute of Mathematical statistics, American Mathematical society providence, Rhode Island, 1986

[49] S.E. Edgell and S.M. Noon, *Effect of violation of normality on the t-test of the correlation coefficient.* Psychological Bulletin, 95, 576-583.

[50] E. Støa, O. Støren, E. Enoksen, F. Ingjer, *Percent Utilization of VO2max at 5-km Competition Velocity Does Not Determine Time Performance at 5 km Among Elite Distance Runners.* J Strength Cond Res. 2010 Apr 9.

[51] D. Costill, H. Thomason, E Roberts, *Fractional utilization of the aerobic capacity during distance running.* Med Sci Sports. 5:248-52, 1973.

[52] Gwowen Shieh, *Estimation of the simple correlation coefficient.* Behaviour Research Methods 2010, 42(4), 906-917, doi:10.3758/BRM.42.4.906.

[53] B. Efron, *Better Bootstrap Confidence Intervals.* Journal of the American Statistical Association Vol.82, No. 397, 171-185, doi:10.2307/2289144, JSTOR 2289144.

[54] Olkin and Pratt, *Unbiased Estimation of Certain Correlation Coefficients*. The Annals of Mathematical Statistics 29(1):201-211. doi:10.1214/aoms/1177706717, JSTOR 2237306.